

Internal Migration in the Survey of Family, Income and Employment (SoFIE)

Ying Li

*School of Mathematics, Statistics and Operations Research
Victoria University of Wellington,
Wellington, New Zealand*

1/11/2010

Key words: Internal Migration Families; Longitudinal Analysis; Longitudinal Survey; SoFIE

Abstract

This thesis studies the short term outcomes in family annual income changes associated with internal migration in New Zealand. This is carried out through analysis of the unit record dataset produced from the Survey of Family, Income and Employment (SoFIE) which is a longitudinal social survey undertaken by Statistics New Zealand starting in October 2002.

Results from data analysis shows us that family migration status and time of measurement have a close relationship with family annual income. Moreover, family structure, average age of adult members in a family, education level of family members, ownership of family residential property and the number of dependent children in a family are all factors related to family annual income.

We apply the General Linear Mixed Model to control these confounding variables and explore the relationship between migration status and annual income of a family. The results from our final model show that changes in the family annual income from before to after their internal migration are different for families. The difference depends on the average age of adult family members. Families with a younger average age benefit from internal migration: their family annual incomes are increased shortly after their movement. In contrast, families with a older average age experience a loss of income.

Acknowledgments

Firstly I would like to thank my supervisor Ivy Liu and co-supervisor Richard Arnold for their continuous support in my Masters program. They taught me how to structure a thesis, set up research questions, explore the relationship between factors in Exploratory Data Analysis, and also how to set up statistical model and analyse outputs. They were always there to listen and to give advice.

A special thanks goes to my manager Andrea Blackburn, without her support in the past three years I couldn't complete my Masters study while working full time.

Let me also say 'thank you' to my colleague Mark Arthur, with his help I finally found a way to generate the longitudinal family ID, and to my colleague Carolyn Fyfe for her proof reading.

I also would like to acknowledge Statistics New Zealand who supplied data from the Survey of Family, Income and Employment.

Most especially, thank you to my family and friends.

Confidentiality statement

To protect the information of individuals, data has been used in accordance with confidentiality requirements, and this thesis has been checked by staff of Statistics New Zealand before it was published. The confidentiality rules applied in this thesis as follows. For graphs, outliers have been removed. For weighted counts, they are rounded to the nearest 1000. For unweighted counts, if greater than or equal to 10 then they are rounded to the nearest 10, zero remains zero, and if greater than zero and less than 10 then they are rounded to 10. Therefore in tables, counts do not necessarily add up to the total, and percentage totals do not necessarily sum to 100%.

Disclaimer

The opinions, findings, recommendations, conclusions and/or even errors expressed in this report are those of the author, not Statistics New Zealand. Access to the SoFIE data was provided by Statistics New Zealand in a secure environment designed to give effect to the confidentiality provisions of the Statistics Act 1975.

Contents

1	Introduction	9
1.1	Previous research on internal migration using longitudinal survey data	10
1.2	Analysis unit and research question	12
1.3	Plan of this thesis	14
2	The survey of Family, Income and Employment	15
2.1	Background	15
2.2	Survey objectives	15
2.3	Survey design	16
2.4	Survey data	17
3	Framework of this research	24
3.1	Definitions and concepts	24
3.1.1	Internal migration	24
3.1.2	Internal migrants and the control group	27
3.1.3	The concept of an intact family	29
3.2	Key variables used from the microdata of the survey	34
3.3	Data derivation and cleaning	36
4	Exploratory Data Analysis	45
4.1	Counts of movers, stayers and intact families	45
4.2	Who are movers?	46
4.3	Primary predictors of family annual income for intact families	50
4.4	Secondary predictors of family annual income for intact families	53
5	Analysis and inference	70
5.1	Introduction analytical methods	70
5.1.1	General Linear Model	70

5.1.2	General Linear Mixed Model	72
5.1.3	Covariance structure for R	74
5.2	Applying the General Linear Mixed model in our case	75
5.2.1	Source of random variation	75
5.2.2	Selecting covariance structure for our data	76
5.2.3	Setting up the General Linear Mixed Model	79
5.2.4	Using the PROC MIXED procedure from SAS	81
5.3	Model selection and analysis	84
5.4	Estimation and inference	87
5.4.1	Estimates on main effects	87
5.4.2	Impacts on log(ai) from the factors of interest: migration and casewave	90
5.4.3	Impacts on log(ai) from migration and casewave for families of different ages	92
5.4.4	Estimating the difference in difference	95
6	Conclusion and future work	99
A	Keys for categorical variables in models	104
B	Initial model for analysis of wave 2 and 4 data	105
C	Final model for analysis of wave 2 and 4 data	120
D	Final model for analysis of wave 1 to 4 data	131
E	Estimation results from offset age 20 in final model wave 2 and 4	141

List of Figures

2.1	Number of OSMs and Cohabs by wave	21
3.1	Movers, Stayers and Others	28
3.2	Diagram of family members' movements	31
3.3	Mover and stayer intact family annual income, wave 2 (income \$000)	41
3.4	Mover and stayer intact family annual income, wave 4 (income \$000)	42
3.5	Logarithm of annual income for mover and stayer intact families, wave 2	43
3.6	Logarithm of annual income for mover and stayer intact families, wave 4	44
4.1	Percentage of adult movers and stayers by age, wave 2 . . .	47
4.2	Family log(ai) changes for mover and stayer intact families (Simple Random Sample of 100 families)	51
4.3	Mean of family log(ai)	52
4.4	Family log(ai) by family type wave 2, where C/O: couple only; C/C: couple with child(ren); OP/C: one parent with child(ren); N/F: not in family nucleus	54
4.5	Changes in the median value of family log(ai) from waves 2 to 4, by wave 2 family type	55
4.6	Family log(ai) by agecat , wave 2	57
4.7	Family log(ai) by agebase in wave 2 and wave 4	58
4.8	Changes in the median value of family log(ai) from wave 2 to 4, by wave 2 age groups (agecat) of the average age of adults in the family	59
4.9	Family log(ai) by highest qualification of adult family members (highqual) wave 2, where no qual: no qualification; sch/voc: school or vocational qualification; univers: university qualification	60

4.10	Changes in median value of family log(ai) from wave 2 to 4, by wave 2 highest qualification of adults (highqual) . . .	61
4.11	Family log(ai) by number of dependent children (depccat), wave 2	62
4.12	Family log(ai) by number of dependent children (depc) for wave 2 and 4	63
4.13	Changes in the median value of family log(ai) from waves 2 to 4, by wave 2 number of dependent children (depccat) .	65
4.14	Family log(ai) by dwelling tenure status wave 2 (tenure) . .	68
4.15	Changes in the median value of family log(ai) from waves 2 to 4, by wave 2 dwelling tenure status (tenure)	69
5.1	Variogram of family log(ai)	78
5.2	Increase rate for family income between wave 4 and wave 2 for mover and stayer families by age point refer to Ratio column in Table 5.6	94
5.3	Summary of Differences in difference in each age points refer to Ratio column in Table 5.7	97
B.1	SAS code for initial model, wave 2 and 4	106
C.1	SAS code for final model, wave 2 and 4	121
D.1	SAS code for final model, wave 1 to 4	132

List of Tables

3.1	Number of intact families from each strategy for movers, stayers	34
4.1	Unweighted counts of OSM movers and stayers in intact families	46
4.2	Adult OSMs in intact families by family types, wave 2 . . .	49
4.3	Highest qualification by migration status, wave 2	49
4.4	Dwelling tenure transition for intact families of movers . . .	67
4.5	Dwelling tenure transition for intact families of stayers . . .	67
5.1	Migration status and wave effects table	83
5.2	Type 3 Test of Fixed Effects, final model for wave 2 and 4 and for waves 1 to 4	86
5.3	Summary of estimates of main effects in final model	89
5.4	Summary of Differences of Least Squares Means of log(ai) between mover and stayer families	90
5.5	Summary of Differences of Least Squares Means of log(ai) between wave 4 and wave 2	91
5.6	Summary of Differences of Least Squares Means of log(ai) between waves 4 and 2, ($\log(E(ai) \text{wave } 4) - \log(E(ai) \text{wave } 2)$), by migration status and age point	92
5.7	Summary of Differences in difference at each age points ($[E(Y_{mw4}) - E(Y_{mw2})] - [E(Y_{sw4}) - E(Y_{sw2})]$)	96
A.1	Keys of categorical variables in models	104
B.1	Model Information, Dimensions, Iteration History	107
B.2	Covariance Parameter Estimates, Fit Statistics	108
B.3	Solution for Fixed Effects (continued to next page)	109
B.3	Solution for Fixed Effects (continued to next page)	110
B.3	Solution for Fixed Effects (continued to next page)	111
B.3	Solution for Fixed Effects (continued to next page)	112

B.3	Solution for Fixed Effects (continued to next page)	113
B.3	Solution for Fixed Effects (continued to next page)	114
B.3	Solution for Fixed Effects (continued to next page)	115
B.3	Solution for Fixed Effects (end)	116
B.4	Type 3 Tests of Fixed Effects (continued to next page)	117
B.4	Type 3 Tests of Fixed Effects (continued to next page)	118
B.4	Type 3 Tests of Fixed Effects (end)	119
C.1	Model Information, Dimensions, Iteration History	122
C.2	Covariance Parameter Estimates, Fit Statistics	123
C.3	Solution for Fixed Effects (continued to next page)	124
C.3	Solution for Fixed Effects (continued to next page)	125
C.3	Solution for Fixed Effects (continued to next page)	126
C.4	Solution for Fixed Effects (end); Type 3 Tests of Fixed Effects (continued to next page)	127
C.5	Type 3 Tests of Fixed Effects (end); Differences of Least Squares Means (continued to next page)	128
C.5	Differences of Least Squares Means (continued to next page)	129
C.6	Differences of Least Squares Means, (end)	130
D.1	Model Information, Dimensions, Iteration History	133
D.2	Covariance Parameter Estimates, Fit Statistics	134
D.3	Solution for Fixed Effects (continued to next page)	135
D.3	Solution for Fixed Effects (continued to next page)	136
D.3	Solution for Fixed Effects (continued to next page)	137
D.3	Solution for Fixed Effects (end)	138
D.4	Type 3 Tests of Fixed Effects (continued to next page)	139
D.4	Type 3 Tests of Fixed Effects (end)	140
E.1	Model Information, Dimensions	142
E.2	Solution for Fixed Effects (continued to next page)	143
E.2	Solution for Fixed Effects (continued to next page)	144
E.2	Solution for Fixed Effects (continued to next page)	145
E.2	Solution for Fixed Effects (end)	146

Chapter 1

Introduction

This thesis studies the effect of internal migration on family annual income. Does the whole family gain financially due to internal migration? If the answer is 'Yes', how much does the family gain compared to families that stay. Internal migration is the phenomenon of people migrating within the same nation. The reasons for migration are many and varied. The migrants may have family, social or cultural reasons, they may move for health or educational purpose, or perhaps for employment. The improvement of income or material circumstances is a common drive for internal migrants. In this thesis we investigate the financial benefits to families of internal migration.

Longitudinal survey data has been used by many researchers to study changes over time. The unique merit of a longitudinal survey is that it is designed to follow a selected sample and interview them repeatedly over time. In contrast, cross-sectional surveys are designed to interview selected samples at specific points in time. Although, some cross-sectional surveys can be conducted repeatedly, a different samples is involved each time.

To infer outcomes regarding internal migration, cross-sectional survey data provides estimates of a sample at a single point in time. Characteristics of the sample changed over time, and these changes cannot be captured by the cross-sectional survey. In contrast, the defining feature of repeated measurements for the same group of individuals over time in a longitudinal survey allows those changes in characteristics to be taken into account. Even if the cross-sectional survey was repeated, as the samples are different each time, therefore it is much less sensitive and powerful compared to a longitudinal survey. Estimates from a longitudinal survey reflect trends

of changes in characteristics over time.

Moreover, confounding from unmeasurable characteristics of an individual can be controlled when using longitudinal survey data as people are different. Differences occur in measurable characteristics such as occupation, education level and working experience, and in unmeasurable characteristics like determination or cooperation. Those characteristics play important roles in impacting their income changes after their internal migration. The unmeasurable factors most likely remain consistent over time for each individual. Longitudinal data controls for variations in these aspects through repeat measurements within the people themselves.

This thesis uses the unit record dataset from the longitudinal survey, 'The Survey of Family, Income and Employment' (SoFIE), to study economic situation changes in terms of the income changes associated with internal migration.

1.1 Previous research on internal migration using longitudinal survey data

Previous research has studied this phenomenon by using longitudinal survey data. Many interesting findings are summarised as follows.

The Panel Study of Income Dynamics (PSID) is a longitudinal survey which started in 1968 with individual samples from 4,800 families in the USA. At its conclusion in 2003, it had collected information from more than 65,000 individuals, spanning as much as 36 years of their lives. Krieg (1997) used microdata from this survey to study the financial returns of internal migrations. In his research he found that migration was an insignificant factor when determining earnings. 'The influence of migration is revealed through the interaction variables.' which shows the relationship between migration and both changes in occupation and changes in employment.

Using data from the National Longitudinal Survey of Youth 1979 (NLSY79)*,

*The NLSY79 started in 1979 with 12,686 young men and women aged 14 to 22 years at the time of their first interview. It provided detailed data of life-course experiences of young people born in the late 1950s and early 1960s, living in the United States.

Yankow (2003) found that the changes in earnings after internal migration are highly related to migrants' education levels. 'Less educated workers do best when changing jobs across labour markets', and 'highly educated workers demonstrate significant extended returns to migration with the bulk of pecuniary rewards accruing with a lag of nearly two years'.

Andrews et al. (2007) examined the labour market returns to internal migration in the United Kingdom by using the data from the first 14 waves of the British Household Panel Survey (BHPS)[†]. This research compared the circumstances before and after migration, and studied labour market activities and earnings over time. They found that migrants who moved for a job experienced a considerable increase in their wages after a lag of about three years. They also found that migration impacts on employment status differently depending on sex, in particular, female migrants who moved for non-employment reasons had a large decrease in employment immediately after their move, while the impact for males was not significant.

Using data from the same survey, Blackburn (2006) stated the similar result, that migration impacts differently depending on sex. He found that wives' earnings fell after a change in location, due to a reduction in working hours. However, for their husbands, the migration event was associated with increased earnings.

Research from Bill & Mitchell (2006) showed that unemployed people are more likely to be involved in repeated moves. They also stated that migration does not play a key role in the transition from unemployment to employment in Australia, when compared with impacts from the employability. They used data from the Household Income and Labour Dynamics Australia study (HILDA), which began in 2001 with 19,914 original sample members.

In conclusion, internal migration has attracted attention from many researchers all over the world. Some of the interesting topics that have already been studied are the internal migrants' financial returns through internal migration, the impact on an individual's labour market activities and so on. However many of previous research explored the impact on individual but not on family. As family is a collective of individuals and a

[†]This survey was initiated in 1991. It reinterviewed 10,300 original individual samples in 5,500 households annually.

component of society, the aim of this thesis is to study the impacts on family income of internal migration in New Zealand, using the unit record dataset of SoFIE.

1.2 Analysis unit and research question

In this section we define the analysis unit of interest. As this thesis uses the SoFIE data, we need to know what collection units exist in this survey and define an analysis unit to meet the purpose of our thesis.

Common forms of collection unit for social surveys are ‘individual’, ‘household’, ‘dwelling’, or ‘family’. Statistics New Zealand defines a ‘household’ as ‘either one person who usually resides alone, or two or more people who usually reside together and share facilities (such as eating facilities, cooking facilities, bathroom and toilet facilities, and a living area), in a private dwelling.’, source of *Household definitions* (2006); while ‘dwelling’ is more about a physical construction. It is ‘any building or structure, or part thereof, that is used (or intended to be used) for the purpose of human habitation. It can be of a permanent or temporary nature and includes structures such as houses, motels, hotels, prisons, motor homes, huts and tents. There can be more than one dwelling within a building, for example an apartment building where each separate apartment or unit is considered a dwelling’, source of *Dwelling definitions* (2006).

Like many longitudinal surveys, the SoFIE is designed to provide information about changes in individual circumstances by re-interviewing the same people over time. The usual unit of collection is an individual person, as individuals are much easier to track than are entire families or households. Family composition often changes as time goes on. For example, adult children may move back to live with their parents, or a couple only family may become a couple with a new born baby. Identifying an unchanged family unit during the longitudinal survey period is difficult, therefore, most of the longitudinal surveys use individuals as their unit of collection.

Although, the migration of a family may be caused by a reason specifically related to a particular family member, the phenomenon happens to the whole family, not just the individual. In this case, not all family mem-

bers are 'active movers', some of them are 'following movers'. For example, take a family consisting of a couple with two children. If the husband found a job in another city and the whole family moved with him, the husband is the 'active mover', while the wife and children are 'following movers'. Internal migration may benefit the husband in terms of his career progression and increased income, but may impact the rest of the family negatively. Some previous studies have treated individuals equally when studying internal migration impacts, others have studied impacts on 'active movers' and 'following movers' separately, but very few have studied this topic at the family level.

It is however important to study the impact of internal migration at the family level. Changes at the family level are an accumulation of the impacts on individual family members. Going back to the example above, suppose after internal migration, income increased for the husband, but decreased for the wife. To study how the family's financial situation has changed after migration, we need to combine the increase and decrease to see the total change in family income. As all family members are living in the same financial environment, they will be similarly influenced by the impact of internal migration. In our example family, if the total impact on family income is positive, then assuming other factors are unchanged, we can say this family is in a better financial situation after their move, and all family members will benefit from this improved situation. In contrast, if the total impact is negative, every member of the family will experience a worse financial situation, even if the husband increases his own income. Therefore, studying the impact of internal migration on a family will also capture the final impact on each of the family members.

The analysis unit used in this thesis is the family nucleus. A family nucleus is defined by Statistics New Zealand as 'a couple, with or without child(ren), or one parent and their child(ren) where the children do not have partners or children of their own living in the same household. Note that the children can be of any age', source of *Family definitions* (2006)

In summary, the research questions of this thesis include: are the changes in a family's annual income (used a measure of a family's financial situation) related to internal migration and which explanatory variables are associated with the changes in a family's annual income?

1.3 Plan of this thesis

Following this general introduction, the next chapter introduces the SoFIE survey and microdata used in this thesis. Chapter 3 gives further detail of the concepts underling the research, and how longitudinal family units are constructed. Exploratory Data Analysis (EDA) is carried out in Chapter 4, where we investigate the relationship between family annual income and other explanatory variables, to identify any potential confounding factors for more complex analysis to follow. A statistical model is set up and results are analysed in Chapter 5. Chapter 6 summarises the key findings of this thesis.

Chapter 2

The survey of Family, Income and Employment

In this chapter, we introduce the Survey of Family, Income and Employment (SoFIE), and outline aspects of the data processing relevant to this thesis. The following contents are mainly drawn from *A Longitudinal Survey of Income, Employment and Family Dynamics, Feasibility Project Final Report* (2001).

2.1 Background

The SoFIE is a longitudinal survey conducted by Statistics New Zealand. It has a fixed sample panel design, which is defined by Schutt (2006) as ‘a type of longitudinal study in which data are collected from the same individuals- the panel- at two or more points in time’. The feasibility study for this longitudinal survey started in 1997 and a final feasibility report was produced in 2001. The survey had been in field since October 2002, and will be finished in 2010. Due to the availability of the data, this thesis uses the unit record data from waves 1, 2, 3 and 4.

2.2 Survey objectives

As the first nationwide longitudinal survey in New Zealand undertaken by Statistics New Zealand, the overall objective of the SoFIE is stated in

A Longitudinal Survey of Income, Employment and Family Dynamics, Feasibility Project Final Report (2001) as 'to provide information about changes over time in the economic well-being of individuals and their families, and about factors influencing those changes'. The longitudinal objectives are divided into four main groups: 'income dynamics (individual, family and household); family formation and dynamics; changes in assets and liabilities and labour market and education experience.'.

2.3 Survey design

Panel length and wave length

The fixed panel for SoFIE is eight years or 'waves'. Longitudinal respondents selected into the panel are asked to recall information from the 12 months prior to the interview date, meaning that the length of each wave is one year. Attempts are made to re-interview respondents as close to 12 months after their initial interview as possible.

Target population

The target population stated in *A Longitudinal Survey of Income, Employment and Family Dynamics, Feasibility Project Final Report* (2001) is 'the usually resident population of New Zealand, living in private dwellings.' This means that the following groups are ineligible for the survey, 'overseas visitors who intend to stay in New Zealand for less than 12 months, non-New Zealand diplomats and diplomatic staff and their dependants, members of non-New Zealand armed forces stationed in New Zealand and their dependants'.

Sample frame

The SoFIE used Statistics New Zealand's existing area-based frame as its sample frame. This is based on a set of small geographic areas, called Primary Sampling Units (PSUs), which contain about 70 dwellings each. PSUs cover all of the North Island, South Island, and Waiheke Island.

Sample selection

There are three stages to the sample selection process, which are a sample of PSUs, a sample of households within selected PSUs and then a selection of all eligible people within the selected households.

Sample size

Approximately 15,100 randomly-selected households were approached to take part in the survey at wave one. In *Technical Notes, Survey of Family, Income and Employment: Wave Four- September 2006* (2008), it states that 'all adults responding at wave one, and children aged under 15 years for which data was collected in wave one, are known as original sample members (OSMs)'. At wave one there were approximately 29,690 OSMs. By wave four, 22,600 of these remained in the survey. At all waves any people living with OSMs are interviewed. However, they do not become panel members. They are interviewed only if they continue to live with an OSM. Those new people who began living with an OSM, either because they joined an OSM's household, or because an OSM moved into their household, are called 'cohabitants' (Cohabs).

2.4 Survey data

SoFIE collects household level data and personal level data and these data can be divided into three types.

The types of data

- Time independent data
Data collected once, as it will not change over time. For example, date of birth.
- Time dependent data: Point-in-time data
Data relates to a single date, usually the interview date. For example, self-rated health status.
- Time dependent data: Spell data
Data relates to a period of time. This may be the annual reference period which is the 12-month period starting from the first day of the allocation month* in previous year to the last day of the month previous to the allocation month in the current year. Also it can be the interview reference period, which is the period from one day after the previous interview date to the current interview date. Or it

*In wave one, interviewing of those respondents selected into the panel was spread across 12 months, and the month of their interview is their allocation month and it is fixed for all eight waves for most of those respondents.

can be any period reported within those reference periods. For example, a labour market activity spell is a period within the interview reference period.

Data collection

The SoFIE data is collected in a face to face interview conducted by an interviewer using a Computer Assisted Interviewing (CAI) system installed on their laptop. The interview consists of two separate parts: the Household Questionnaire (HQ), and the Personal Questionnaire (PQ). In each wave the HQ is completed by an adult household member, and then all eligible adults within the household are asked to complete a PQ. The modules included in the HQ and PQ are as follows:

- Household Questionnaire (HQ) contains two modules:
 - Household (for example, household type and family type)
 - Standard of living (for example, type of housing and appliances owned)
- Personal Questionnaire (PQ) contains eight standard modules:
 - Demographics (for example, age and country of birth)
 - Child (demographics about any child(ren) under 15 years residing in the household, answered by a nominated adult)
 - Labour market history (for example, age at first paid job)
 - Education (for example, highest qualification)
 - Family (for example, existence of partners and/or children)
 - Labour market (current activity, for example, details of employment)

- Income (for example, from paid work, superannuation, or government payments)
- Contact (respondent's contact details for ease of follow-up).

In waves 3, 5 and 7, the Health module is included in the PQ; in waves 2, 4, 6 and 8, the Assets and Liabilities module is contained in the PQ. Cohabs are not asked the Family module or Health module. This is stated in *Technical Notes, Survey of Family, Income and Employment: Wave Four- September 2006* (2008).

Data processing

In the SoFIE, the collected data is re-organised, edited and coded before it goes into the User System. The User System is to produce the unit record dataset through imputation, derivation and weighting and estimation processes.

- **Imputation**
Some key variables missing from both item non-response and wave non-response will be imputed. Item non-response is where respondents refuse or don't know the answer for an item in questionnaire. Wave non-response is where respondents miss an entire interview for a wave. Retrospective data are used for imputation in that case. If OSM respondents missed one wave, then they will be asked some key information of their missing wave in their immediate following wave. The imputation process populates that information to their missing wave.
- **Derivation**
To provide comprehensive output, the SoFIE unit record dataset provides a great number of derived variables by the Derivation process. For example, personal annual income is a derived variable which is the summation of personal income from all sources in a particular annual reference period.
- **Weighting and Estimation**
The SoFIE wave 1-4 unit record dataset contains longitudinal and

cross-sectional weights. This section just gives a brief description of weights, detailed calculation of weights is not provided as this thesis is focusing on unweighted analysis. Longitudinal weights reflect the population at the time of sample selection, the 2002 New Zealand population, and cross-sectional weights reflect the population at the end of each wave.

If N_t is the population of New Zealand at wave t , then at all waves t ,

$$\sum_{i: \text{responding OSMs}} W_{t,i}^L = N_1$$

$$\sum_{i: \text{all respondents}} W_{t,i}^C = N_t$$

where $W_{t,i}^L$ = longitudinal weights for respondent i , i for OSMs only; and $W_{t,i}^C$ = Cross-sectional weight for respondent i which includes OSMs and Cohabs.

Figure 2.1 shows the predicted changes in the number of OSMs and Cohabs over the eight waves of the survey. In wave 1, all of the eligible responding samples members are OSMs. The number of OSMs is represented by the red bar in Figure 2.1 (page 21). They have been assigned longitudinal weights, which sum up to the 2002 New Zealand population. As there were no Cohabs in wave one, the total of longitudinal weights equals the total of cross-sectional weights. From wave 2 onwards, the number of OSMs represented by blue bars decreased over time as sample attrition. The attrition may due to natural deaths, loss to tracking, moving overseas or to an institution. The remaining number of OSMs decrease by wave, 89% remain in wave two, 82% in wave three and 76% in wave four reported by *Technical Notes, Survey of Family, Income and Employment: Wave Four- September 2006* (2008). To reflect the same population in wave 1, the longitudinal weights re-assigned to the remaining OSMs are slightly higher in each of the following waves. Meanwhile, the number of Cohabs may increase over time, as OSM households break up and form new households which introduce new Cohabs into the survey. This is represented by the yellow bars in Figure 2.1 (page 21). From wave 2 onwards, the total number of cross-sectional sample members is the sum of

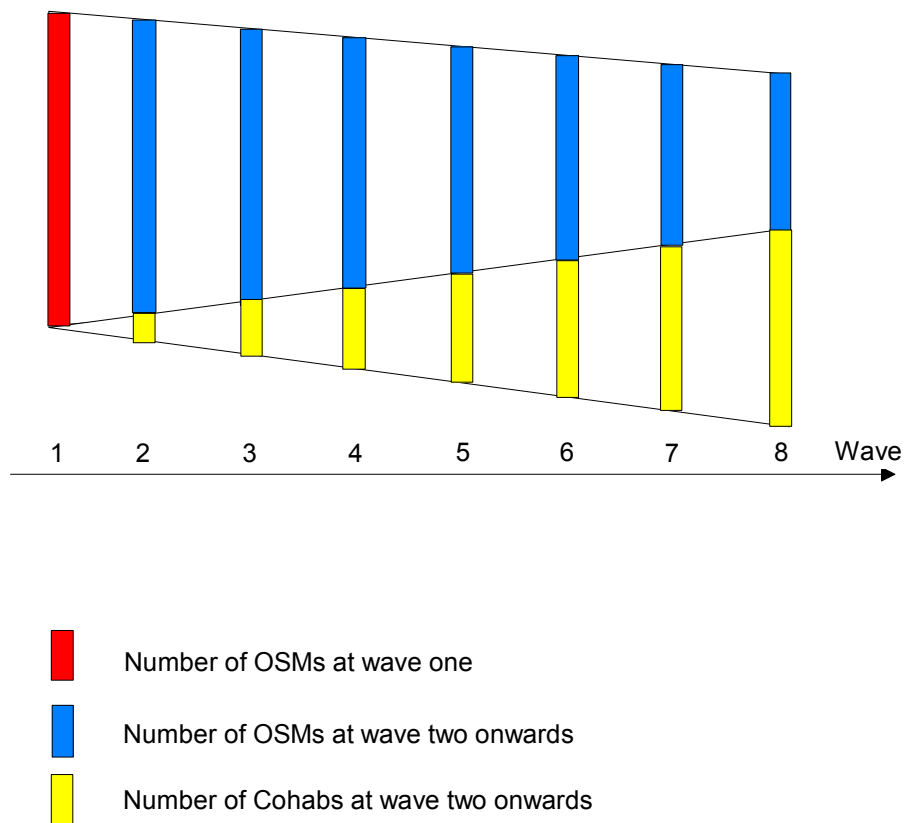


Figure 2.1: Number of OSMs and Cohabs by wave

the OSMs and Cohabs in that wave. Cross-sectional weights are calculated and assigned to both of them. To reflect population growth over time, the sum of the cross-sectional weights increases with each wave.

In the SoFIE unit record dataset, longitudinal weights have not been provided for the unit of analysis in this thesis, families. In this thesis, we are interested in regression relationships among exploratory variables and income changes within individual family, and are not aiming to model estimates of population means or totals, therefore only unweighted data has been used here and weighting calculations will not be discussed in further detail.

Attrition and response rate

The attrition of longitudinal sample across waves is a common issue for fixed sample panel longitudinal surveys. To keep the attrition rate to a minimum, the SoFIE is collected using the following tracking rules: OSMs should be tracked unless they have missed two or more consecutive waves of interviewing; if an OSM misses one wave of interviewing, they will continue to be tracked. If they respond in the following wave, they will be asked to supply information relating to the wave they missed. The missed wave is then known as a 'sandwich missing wave'. The OSM response rates of wave 2, 3 and 4 were 89%, 82% and 76% respectively.

Data limitations and challenges for this study

Due to the longitudinal nature of the survey, and the richness of the data collected, the SoFIE is an unique resource from which to study changes in people's lives over time. However there are still some limitations.

The most recent release of microdata from the SoFIE only contains information from waves 1 to 4, thereby limiting this thesis to the study of the short-term impacts of internal migration only.

As the analysis unit of the survey is the individual, the collection aims to re-interview individual OSMs. But in order to study the impact of internal migration on family, the concept of a longitudinal family is required. However, the fact that family composition changes over time, makes defining a longitudinal family a challenge. For such an analysis to be feasible, an associated longitudinal family identifier must be created. In addition, the survey data is collected and structured on an individual basis. Thus, another challenge for this thesis was the reorganisation of the data, and the

recalculation of derived variables, at the family level.

Confidentiality rules applied in this thesis

As data confidentiality is a concern, this thesis has been checked by staff of Statistics New Zealand before it was published. The confidentiality rules applied in this thesis as follows. For graphs, outliers have been removed. For weighted counts, they are rounded to the nearest 1000. For unweighted counts, if greater than or equal to 10 then they are rounded to the nearest 10, zero remains zero, and if greater than zero and less than 10 then they are rounded to 10. Therefore in tables, counts do not necessarily add up to the total, and percentage totals do not necessarily sum to 100%.

Chapter 3

Framework of this research

In this chapter, we define key concepts, identify populations of interest, select variables from the microdata, and derive variables for use in this analysis.

3.1 Definitions and concepts

As this thesis is interested in studying the impacts and outcomes of internal migration on the economic situation of a family, the first key concept that needs to be defined is internal migration. We will then describe who the people involved in this event are, and explain what a family is in a longitudinal context, for the purpose of this thesis.

3.1.1 Internal migration

There are two aspects which must be considered when defining internal migration - distance and time. Both of these are discussed below.

Distance

In previous research, the actual distance people moved has been used to define internal migration. For instance, when using the BHPS to study the impact of internal migration within Britain on married couple's earnings Blackburn (2006) defined a change in location as a move of more than 50 kilometres. In Australia, Bill & Mitchell (2006) calculated the distance of a move by means of great circle distance between postcodes reported in HILDA. That investigated the relationship between migration and labour

market outcomes. In other studies geographic boundaries are used. For example, Andrews et al. (2007) defined internal migration using movements across the eleven standard statistical regions of the UK to study the employment and earnings of migrants in UK.

For confidentiality reasons, the only geographic information available in the SoFIE microdata is Major Region. Major Regions are the aggregated standard regional council areas. It categorises New Zealand into six regions: Auckland, Waikato, Wellington, Rest of North Island, Canterbury, and Rest of South Island. An indicator of a change in Territorial Authority (TA)*. TA is available as additional data to the SoFIE wave 1-4 microdata and it has been produced for wave 2 onwards. However, the TAs involved are not specified. Apart from this, lower level geographic information, such as meshblock[†] and household address, are not provided. This limitation means this thesis can not define internal migration by the actual distance moved.

Major Regions are too large to be used to define migration, since very long distance moves are possible while still remaining within a Major Region. Moreover, short distance moves between urban and rural areas are not detectable. Another reason for not using Major Region to define internal migration is that it may not provide a sufficient sample to study. As many researchers have shown, the majority of movements are over small distances. Bill & Mitchell (2006) found that in waves 2, 3 and 4 of HILDA, one-third of movers stayed inside their own postcode, and over half moved nine kilometres or less. Movements between Major Regions may only capture a small proportion of internal migration, due to the distances covered. Therefore, Major Region is not suitable to define internal migration in this thesis.

As mentioned above, an indicator of a change in TA is available. It indicates if the respondent has moved TAs between waves. As TAs are smaller units than Major Regions, using the indicator of a change in TA will capture a greater number of samples than using Major Region would. In addition, even we realised that some movements between TAs may just change residence to an adjacent TA, moving between TAs is still a more

*A Territorial Authority is defined under the Local Government Act 2002 as a city council or district council. There are 73 territorial authorities, comprising of 15 cities and 58 districts, source of *Territorial Authority* (2002)

[†]The meshblock is the smallest geographic unit for which statistical data is collected and processed by Statistics New Zealand. *Meshblocks* (2001).

significant event when considering changes in people's lives, than is moving between smaller level aggregations, such as Area unit[‡] or Meshblocks. This thesis therefore uses movements between TAs to define internal migration instead of actual distance.

Time

The residential address of the dwelling is point-in-time data, which is collected each wave at the Household Enumeration Date (HED)[§]. This means that the respondent lives at the reported address at the HED. However, this respondent could have lived at another address, or addresses, between the previous wave's HED and the current HED. Similarly, the indicator of a change in TA will indicate if the respondent lived in a different TA, as at the HED. But it will not necessarily reveal if they lived in the same TA throughout the entire period. It is assumed here that if the indicator of a change in TA is 'no', then the respondent has lived in the same TA from the previous HED to the current HED, although they may in fact have lived in different TAs between these two dates.

Previous research shows that the outcomes associated with internal migration are highly dependent on how long it has been since the movement. Borjas, Bronars and Trejo (1992) concluded that internal migrants initially earn less than natives, but that difference disappears within a few years. Other research has shown that the immediate financial loss depends on the migrants' level of education and occupation. For instance, Yankow (2003) found 'contemporaneous returns are found only for migrants with twelve or less years of completed schooling, in contrast, highly educated migrants receive significant extended returns to migration with a lag of nearly two years'. This shows that the length of time since the migration plays an important role in determining outcomes of internal migration.

To study the outcomes and associated factors of internal migration, this thesis compares the financial situation of internal migrants before and after movements. As data is only available for waves 1-4, this research is only able to study the short-term outcomes associated with internal mi-

[‡]Area units are aggregations of Meshblocks. They are non-administrative areas that are in between Meshblocks and Territorial Authorities in size, *Area unit* (2006).

























































[§]The household enumeration date (HED) is the date when all eligible members of a household have been identified and recorded in the household questionnaire.' (Statistics New Zealand, March 2008)


gration. In order to make the situation before and after movement comparable, it is necessary to examine a consistent time period. The model set up in this thesis will therefore look at one movement during these four waves only. The period of time selected is between wave 2 and wave 3 HEDs. In this model, the time specific aspect of internal migration is a changing TA indicator. That is, the respondent has a 'no' for wave 2 and wave 4, but a 'yes' for wave 3. In other words, the respondent reported staying in the same TA in waves 1 and 2, and in another consistent TA in waves 3 and 4, with a move between TAs occurring in wave 3 only.


3.1.2 Internal migrants and the control group

In wave 1 of the SoFIE wave 1 to 4 micro dataset there are approximately 29,690 OSMs and 21,730 of these provided complete responses in all four waves. This includes for both imputed values and refusals or 'don't know' responses to some questions. According to the definition of internal migration specified, we identify internal migrants from the SoFIE waves 1-4 microdata. Figure 3.1 (page 28) presents different scenarios of movement between TAs across waves. There is one group of movement we are interested in, we call them movers and also the control group we call them stayers. Note that rounding has been applied on those counts, therefore counts do not necessarily add up to the total.

- **Movers**
Movers are OSMs who have only changed their reported residence address between TAs at the household enumeration date of wave 3 during the first four waves' interviews. This is shown in Figure 3.1 as the first scenario. Under this definition, approximately 770 OSM respondents have been categorised as movers.
- **Stayers**
Stayers are OSMs who have not reported a change in their residence address that equates to a move between TAs, during the first four waves of interviewing. This is shown in the second scenario of Figure 3.1. Under this definition, 18,980 OSM respondents are cate-

	wave1	wave2	wave3	wave4	Frequency	Percentage			
Movers								770	3.55%
Stayers								18980	87.47%
Others_1								800	3.69%
Others_2								750	3.46%
Others_3								160	0.74%
Others_4								110	0.51%
Others_5								150	0.69%
Others_6								40	0.18%

 Territorial Authority

 Stay in the same Territorial Authority


 Move between Territorial Authorities

Figure 3.1: Movers, Stayers and Others

gorised into this group as stayers.

- Others

OSMs whose movements follow the other 6 scenarios presented in Figure 3.1 are categorised as others. This group is not of interest in this thesis.

The reason for only using movers and stayers in this thesis is that the income for the previous annual reference period is only reliable for a person who has not moved. We require a relatively stable situation in order to look at the impact of internal migration on income. That means we can calculate an income associated with a particular location using the previous annual income only if the respondent has not moved in that period. Thus, movers' incomes are calculated as wave two reported income (pre move) and wave four reported income (post move). For our main analysis, we do not include wave one or wave three income, because it is not known if the person had moved prior to wave one, and their income is disrupted by internal migration between wave two and wave three. Stayers have no disruptions when calculating annual income.

3.1.3 The concept of an intact family

As discussed in Chapter 1, the family nucleus is the analysis unit of this thesis. However, the members in a family nucleus may change over the survey period. Figure 3.2 (page 31) presents examples of family members' movement between waves. Suppose that in wave t there is a family with a father 'A', a mother 'B' and a child 'C', who are all OSMs. The members of this family may remain unchanged from wave t to wave $t + 1$. This is shown in scenario 1. In scenario 2, the family is split. Father 'A' and mother 'B' still lived together, but child 'C' moved out. This is an example of a young adult leaving their parents household to live alone. Alternatively, additional family members may join together in wave $t + 1$. An example of this is shown in scenario 3, where OSM 'D' moves back to live with 'A', 'B' and 'C', instead of living by herself. Scenario 4 shows an example of a new Cohab joining the OSMs' family. The situation presented here may be that this family has a newborn baby in wave $t + 1$. The last scenario shows the common situation where an OSM misses their interview in wave $t + 1$. This OSM may be tracked in wave $t + 2$ or they may

leave the survey permanently due to an event such as death or emigration. If that is the case, they become the attrition of the sample.

A change in family composition may result in a different family financial situation. To study associated short term outcomes from internal migration for families, factors such as compositions which impact on the family financial situation should be controlled for and remain unchanged. Families who only change their internal migration status, while keeping all of the other factors relating to their family financial situation unchanged, are the families we are interested in. We have called this group **intact families**. Based on this definition, we need to identify the possible combination of factors impacting on the family financial situation, and then work out how to control for them. Four possible strategies can be used to achieve this, which are described below.

(a) Families with the same family type

Previous research by Statistics New Zealand shows that family type[¶] has a close relationship with financial situation of a family. In the report of '2001 Census of Population and Dwellings: Incomes', it classifies family types as 'couple only', 'couple with child(ren)' and 'one parent with child(ren)', and states that

A comparison of family income by family type shows that half of one-parent with children families received \$20,000 or less compared with 5.6 percent of couples with children and 13.7 percent of couples without children. Only 11.9 percent of one-parent with children families received more than \$50,000, compared with 44.7 percent of couples without children and 61.1 percent of couples with children *2001 Census of Population and Dwellings: Incomes* (2001).

We assume that if an OSM has the same family type across all four waves, then they are in an intact family, and can be used in the analysis of internal migration. It is important to recognise here that under this definition the family members could be different individuals, even though the family type remains the same.

[¶]The derived variable that classifies family nuclei according to the presence or absence of couples, parents and children, source of *Family definitions* (2006).

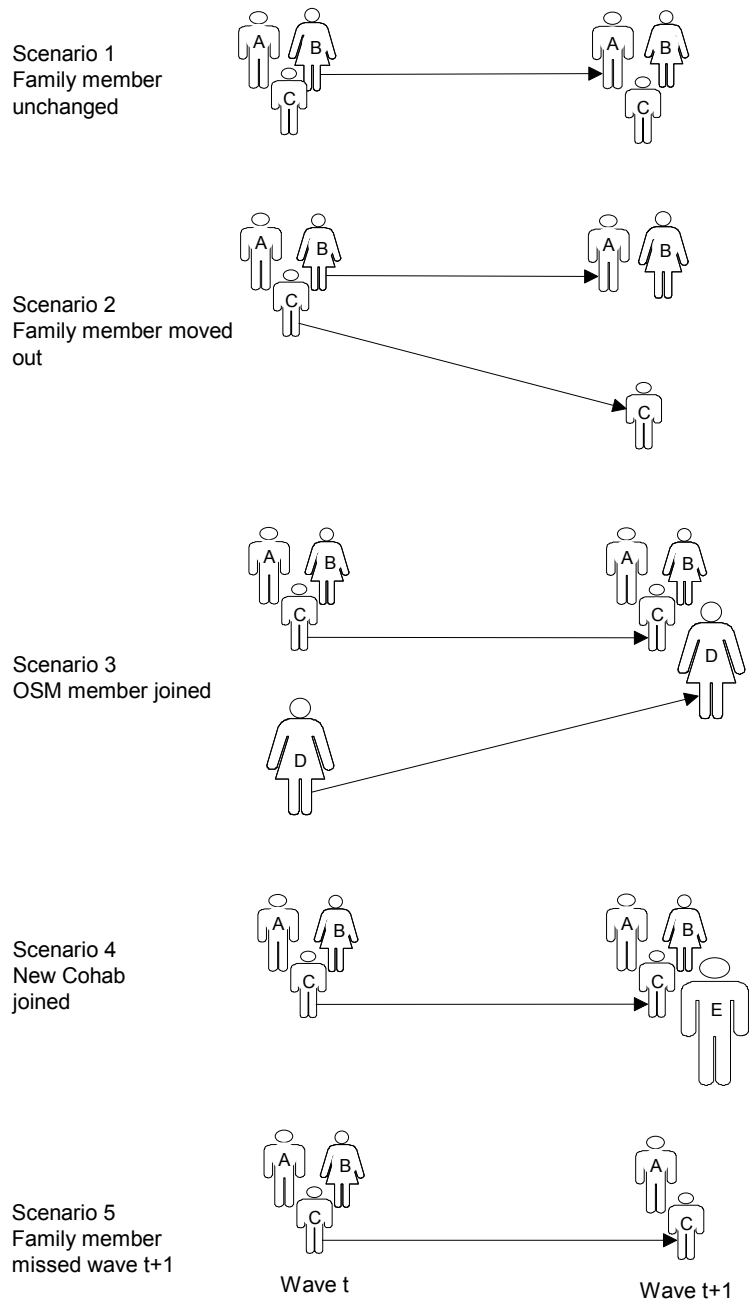


Figure 3.2: Diagram of family members' movements

To identify intact families with the same family type across four waves, we first sort adult family members by their personal identifiers in numerical order, then select the first OSM adult from the family at wave one to be the key adult member^{||} The wave one family type is then attached to the personal identifier of the key adult member to form a family ID. We then use the same method to form the family ID for this key adult member in their following waves. Finally, any family ID which remains unchanged for all four waves identifies an intact family with the same family type.

(b) Families with the same number of adult members

Although family type is closely related to the family financial situation, maintaining the same family type across four waves, does not ensure that the family fits our definition of intact family. For example, even if one or more children join a family with the family type, 'Couple with child(ren)', the family type will not change. The children can be of any age, as this family type includes both a couple with dependent child(ren), and a couple with adult child(ren). Moreover, there is no restriction on the number of adult children that may be included here. Both the number of children in a family and the dependency status of these children may impact on the family financial situation dramatically, merely keeping families where the family type remains unchanged is not enough to identify an intact family.**

As adults are normally the main participants in the labour market and providers of the main source of family income, we assume that if the number of adult members in a family doesn't change, then the family is intact.

To identify intact families with same number of adult members across four waves for an OSM, we use the same strategy as in section (a). The only modification required here is to attach the number of adult members to the personal identifier of the key adult family member instead of family type. This forms a family ID. The family IDs which remain unchanged for all four waves are the longitudinal family IDs for intact families with same number of adult family members.

^{||}the survey is not designed to track OSM children. Only OSM adults are tracked.

^{**}the definition of an adult in this survey is a respondent who is 15 years old or over as at the household enumeration date.

(c) Families with the same adult members

As different adults may contribute to the family economic situation differently, we assume that a family with the same adult members is more likely to keep their financial situation intact. In order to make further improvements to the model, we have included the criterion that the family must contain the same adult members to be considered intact. A similar strategy to that presented in section (a) was used to achieve this. The only difference was that the personal identifiers of the other adult family members' were attached to the personal identifier of the key adult family member in order to form the family ID. As above, any family ID which remains unchanged for all four waves identifies the intact families with same adult members.

(d) Families with the same members including adults and children

In this method, having the same family members including adults and children is the condition to identify intact families. Again, a similar strategy to that presented in (a) was used to achieve this. The only difference in the method used was that the personal identifiers of all the other family members were attached to the key adult member's personal identifier to form the family ID. Again, any family ID which remains unchanged for all four waves identifies the intact families with the same family members.

Based on these four strategies, we calculate the number of intact families of movers and stayers and do a comparison. Note that in this table, and all the subsequent tables, confidentiality rules are applied so counts are rounded to the nearest 10.

Table 3.1 (page 34) shows the number of intact families, calculated using these four strategies, for movers and stayers respectively. It can be seen that the 'same family type' strategy gives the most intact families and the 'same family member' strategy gives the least intact families. To maintain as many intact families as possible to ensure a reasonable sample size to work with, and also to select families to satisfy our definition of intact family which is families who only change their internal migration status while keeping all of other factors relating to family financial situation unchanged, we have chosen to use the third strategy, 'same adult member'.

Table 3.1: Number of intact families from each strategy for movers, stayers

Method	Movers	Stayers
Same family type	200	5910
Same number of adult member	190	5750
Same adult member	190	5700
Same family member	170	5150

This gives us 190 intact families for movers and 5,700 for stayers. Note, there are less than ten families which have not been considered intact families, because they contain children who turn 15 years in one of the three waves after wave 1. This is due to the fact that the number of adult members in the family has changed.

3.2 Key variables used from the microdata of the survey

This section introduces the key data we have used directly from the survey, as well as the source data used for the data derivations produced in this thesis. Descriptions here are mostly drawn from the data dictionary of the SoFIE unit record dataset provided by Statistics New Zealand.

- **PersonLT**- longitudinal personal ID
A randomised number assigned to uniquely identify a person over each wave. For each wave the person has the same identifier.
- **casewave**- Wave
Wave number.
- **AgeAtHed**- age at the Household Enumeration Date (HED)
A derived variable provided in the SoFIE data. This is the age of the respondent, as at the HED. This is based on the date of birth the respondent provided in the demographic module of their PQ.

- **Sex-** sex indicator
Sex of the respondent, as at the HED. This is point-in-time data that is collected at each wave, in the demographic module of the respondent's PQ.
- **Adult-** adult indicator
A derived variable to indicate whether the person is an adult (ie is aged 15 years or more) as at the HED.
- **FamilyNucleus-** family nucleus ID
A unique number to distinguish a family nucleus from another family nucleus in the same household.
- **FamilyStd2-** family type
The Statistics New Zealand Standard Family Type (Level 2) for the person. The categories are: couple only, couple with dependent child(ren) only, couple with dependent and adult child(ren), couple with adult child(ren) only, couple with some or all child(ren) of unknown dependency, one parent with dependent child(ren) only, one parent with dependent and adult child(ren), one parent with adult child(ren) only, one parent with some or all child(ren) of unknown dependency, not in a family nucleus, family type unidentifiable.
- **Childcntdependent-** number of dependent children
This is a derived variable provided in the SoFIE data. It calculates the number of people in the family nucleus flagged as being a dependent child. The dependent children are a.) aged less than 15 as at HED or b.) aged between 15 and 17 years, and not in full-time employment, as at HED.
- **DwellTenure-** dwelling tenure status
This is derived data provided in the SoFIE dataset. It describes the dwelling tenure status of the household. The categories are: owned with mortgage, owned without mortgage, occupied rent-free, rented or leased, dwelling tenure unspecified.
- **Qual-** personal qualification
The post-school qualification code as per the current standard classification for the SoFIE.
- **PersIncAnn-** personal annual income
This is the derived variable of the total income the person received from all sources over the annual reference period.

- **ChangingTAInd**- changing TA indicator
This is a variable derived to indicate if the respondent is living in a different Territorial Authority to the previous wave's interview. This variable is in addition to the wave 1-4 dataset and derived separately by Statistics NZ.

3.3 Data derivation and cleaning

In this section we describe the derivation and data cleaning that are necessary to prepare the data for the EDA presented in next chapter.

- **id**- intact family longitudinal ID
This variable is derived for this thesis to uniquely identify an intact family over each wave. An intact family has the same identifier for all 4 waves.
As discussed above, this thesis uses the same adult members from the family nucleus to define the intact family across all four waves. The steps followed to derive the intact family longitudinal ID are:
 1. Sort the **PersonLT** of adults in each family nucleus in numerical order for all four waves and use the first adult OSM's **PersonLT** to be the key person's identifier.
 2. Create the family **ID** using the key person's **PersonLT** along with each adult member's **PersonLT** concatenated into a long string in the sorted order of step 1. For example, the key person's **PersonLT**=12, and the adult member's **PersonLT**=34 and 56, then the family **ID**=123456. (Note that in the SoFIE all **PersonLT**s have the same number of digits.)
 3. Calculate the frequency of each family ID across all four waves.
 4. Identify longitudinal intact family IDs as those with a frequency of four. Family IDs with frequency counts less than 4 are not intact families.

To identify the cross-sectional information of an intact family, only the combination of the longitudinal family ID and the **casewave** are needed.

- **migration** - family migration status

This is derived from the **ChangingTAInd** and **casewave** number, and is based on the definition of internal migration given above. If **ChangingTAInd** is 'yes' for **casewave** 3 and 'no' for **casewave** 2 and 4, then the migration status is 'mover'; if **ChangingTAInd** is 'no' for **casewave** 2, 3 and 4 then the migration status is 'stayer'. In all other situations the migration status is 'other'. This thesis studies the change of family annual income for 'mover' and 'stayer' groups. The 'stayer' group is used as the control group.

- **familytype**- family type

This variable is derived from **FamliyStd2**. To address confidentiality concerns, some categories of **FamilyStd2** have been collapsed to derive **familytype** to make sure that the number of samples in each category is big enough. The method used to derive this variable is:

- If **FamilyStd2** is 'Couple only' then the **familytype** is 'Couple only';
- If **FamilyStd2** is one of the following, then the **familytype** is 'Couple with child(ren)'.
 - * 'Couple with dependent child(ren) only',
 - * 'Couple with dependent and adult child(ren)'
 - * 'Couple with adult child(ren) only'
 - * 'Couple with some or all child(ren) of unknown dependency';
- If **FamilyStd2** is one of the following, then the **familytype** is 'One parent with dependent child(ren)'.

- * 'One parent with dependent child(ren) only'
- * 'One parent with dependent and adult child(ren)'
- * 'One parent with adult child(ren) only'
- * 'One parent with some or all child(ren) of unknown dependency';

– If **FamilyStd2** is 'Not in a family nucleus' then the **familytype** is 'Not living with a family'. This includes respondents living alone or living with others but not family members.

- **hqual**- personal highest qualification
It is a derived variable which gives the highest qualification of the individual, as of the interview date for each wave. The categories of this variable are grouped are: no qualification, school or vocational qualification, and university qualification.
- **highqual**- family highest qualification
It is the highest qualification held within an intact family for each wave. It has been derived for this thesis, by taking the highest value of **hqual** held by an adult family member in an intact family. The categories of this variable are: no qualification, school or vocational qualification, and university qualification.
- **agebase**- family average age of adult in wave 1
This variable is derived from **AgeAtHed** of the adults in the family, to show the average age of all the adult members of the intact family.
- **agecat**- family average age of adult in wave 1 grouped into 4 levels:
 - '15-25 years',
 - '26-45 years',

- ‘46-65 years’
 - ‘66 years and over’.
- **depc**- the number of dependent children in a family
This variable equals **Childcntdependent**, giving the number of dependent children in a family.
- **depccat**- the number of dependent children in a family grouped into 4 levels:
 - ‘none’,
 - ‘one’,
 - ‘two or three’
 - ‘four or more’.
- **tenure**- family dwelling tenure
Due to confidentiality concerns, this variable is grouped into categories in **DwellTenure** as below:
 - If **DwellTenure** is ‘Owned with mortgage’ or ‘Owned without mortgage’, then **tenure** is ‘Owned type’;
 - If **DwellTenure** is ‘Occupied rent-free’ or ‘Rented or leased’, then **tenure** is ‘Rented type’.
- **ai**- family annual income
Family annual income (**ai**) is an important measurement of the family financial situation. It is derived by adding up the personal annual incomes of all adult family members in an intact family. Figure

3.3 (page 41) and Figure 3.4 (page 42) show the distributions of family annual income for mover and stayer intact families in wave 2 and wave 4. The X-axis represents the **ai** in thousands of dollars and the Y-axis represents the percentage in each class.

It can be seen that the histograms in Figure 3.3 (page 41) and Figure 3.4 (page 42) are all right skewed, this means that the majority of intact families have an annual income between 20,000 to 60,000 dollars. Extreme values are below or equal to zero, or greater than 160,000. As the family annual income is derived by summarising the **PersIncAnn** of all the adult members in the family, the zero or negative annual income may be due to a zero or negative **PersIncAnn**. If a respondent reports their self-employment income as a loss, or zero income from their paid work or self-employment, then they will have zero or negative **PersIncAnn**. For example, if **PersIncAnn**=\$10,000 for adult A, and **PersIncAnn**=\$-10,000 for adult B, then the total family annual income will be zero dollars. On the other hand, data collection errors or processing errors will also result in extreme values in both ends.

- **log(ai)**- the logarithm (based on e) of family annual income
As the proportion of extreme values in family annual income is small (approximately 1% of values are extremely low and 1% are extremely high), we have taken the logarithm of family annual income with a threshold to remove zero and negative values. The data transformation we performed was

$$\log(ai) = \log(\max(1, 1 + ai))$$

Figure 3.5 (page 43), and Figure 3.6 show the logarithm of family annual income for mover and stayer intact families by wave. From Figure 3.5, and Figure 3.6 (page 44) we can see that the logarithm of family annual income is more symmetrically distributed, therefore, this thesis uses **log(ai)**. Note that for the presentation of each graph, the low end is not included, but it is included into the statistical modelling in Chapter 5 of analysis and inference.

Figure 3.3: Mover and stayer intact family annual income, wave 2 (income \$000)

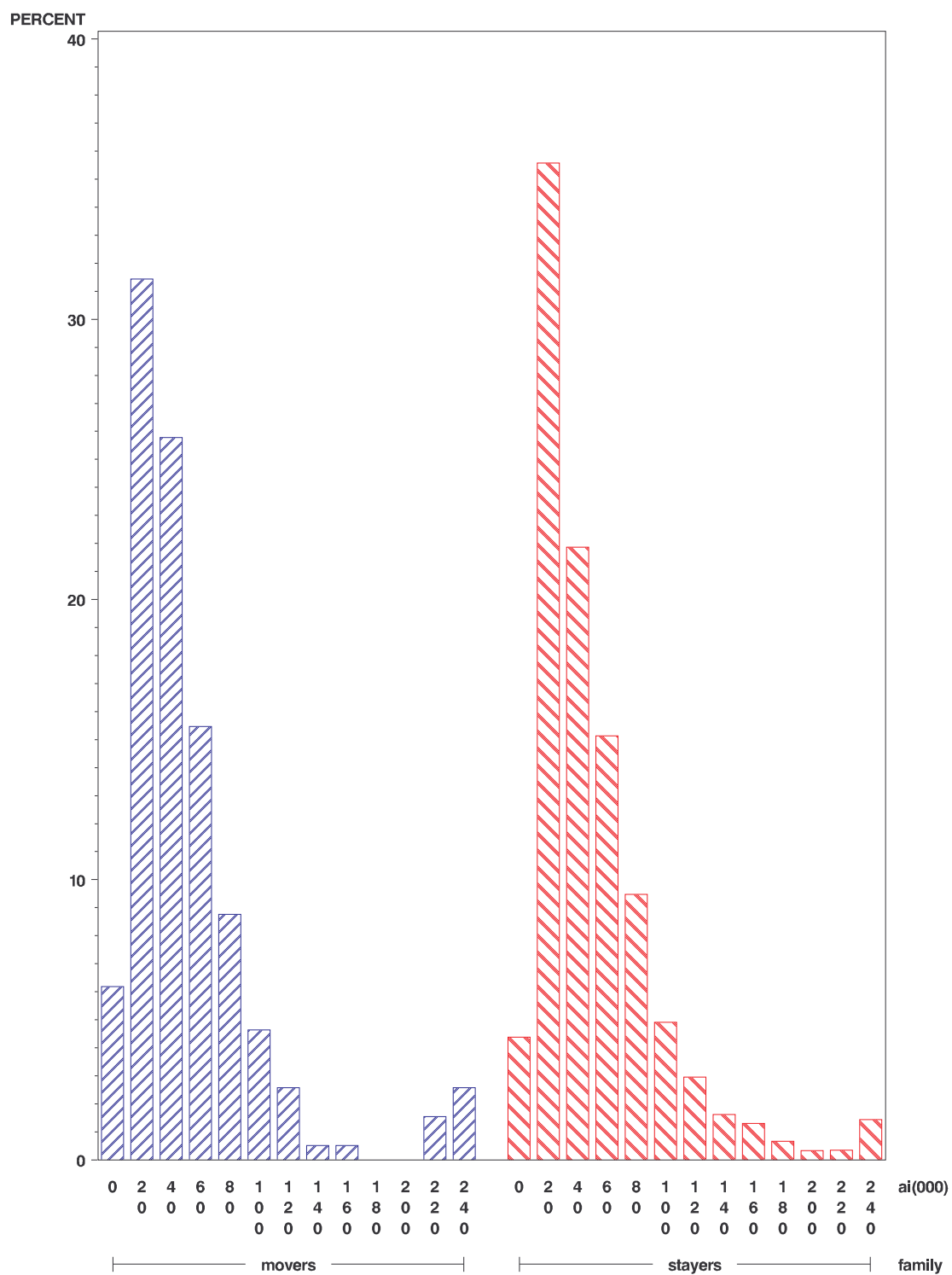
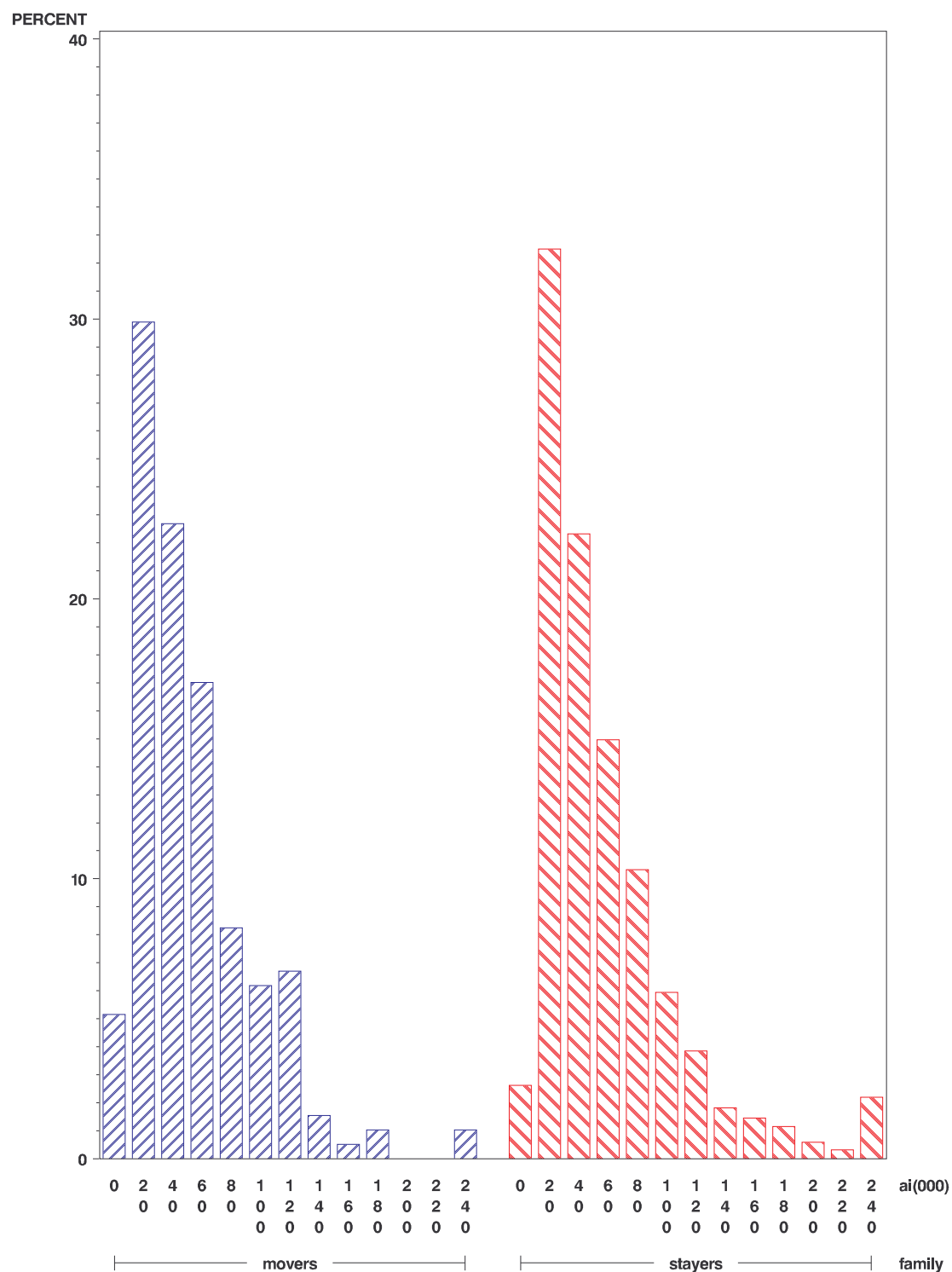


Figure 3.4: Mover and stayer intact family annual income, wave 4 (income \$000)



lies, wave 2

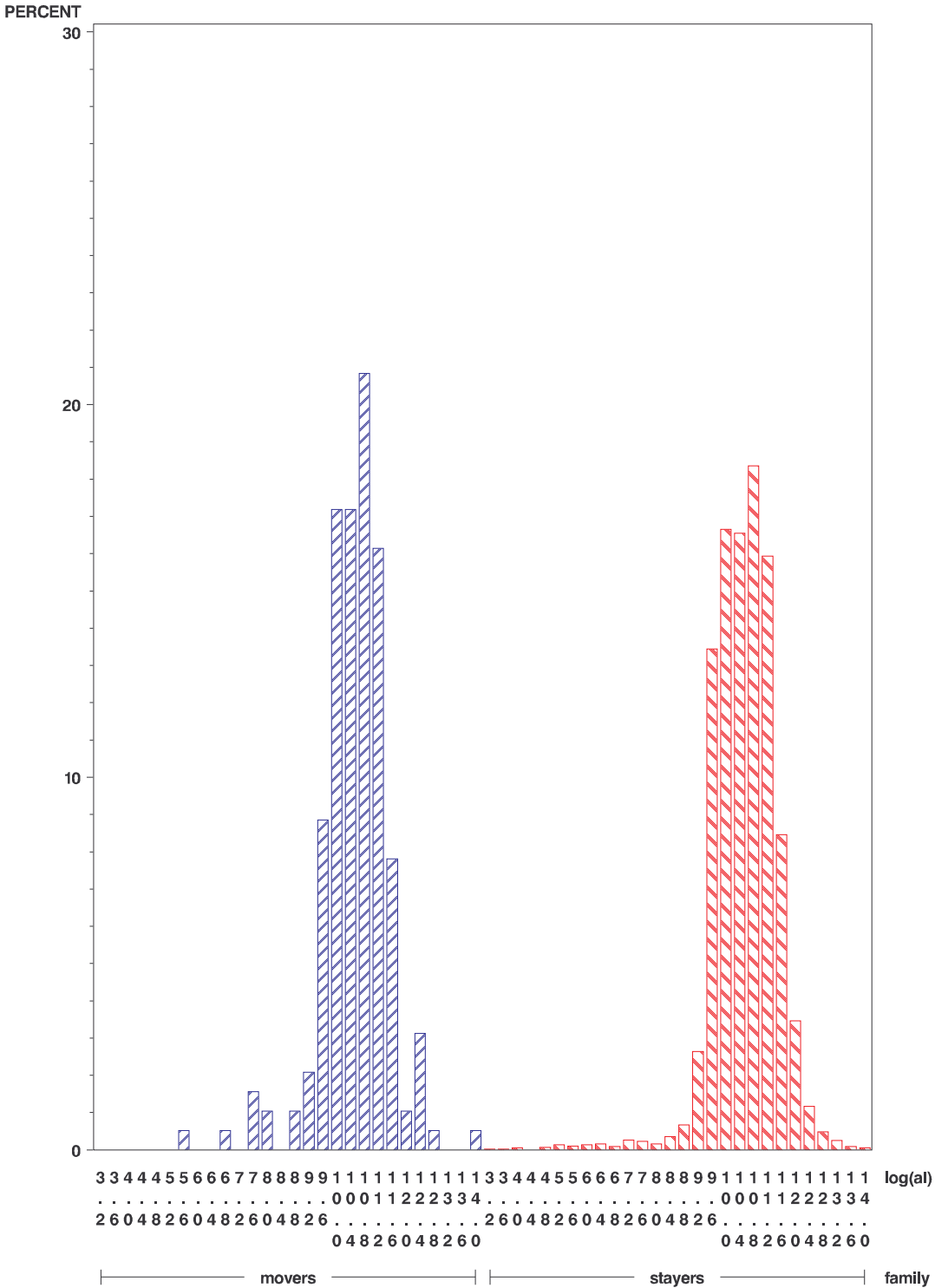
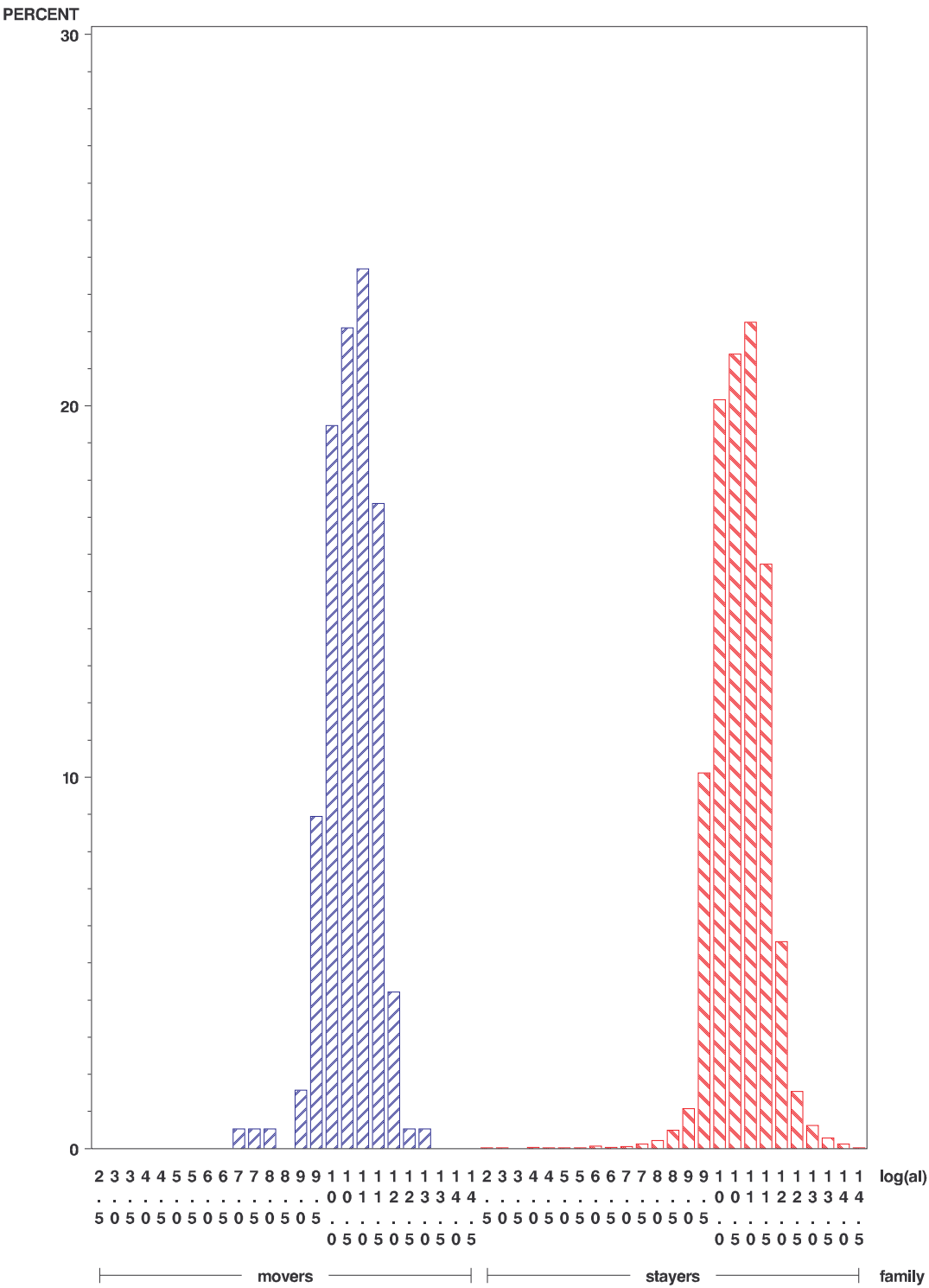


Figure 3.6: Logarithm of annual income for mover and stayer intact families, wave 4



Chapter 4

Exploratory Data Analysis

In this chapter we profile our population of interest and then present Exploratory Data Analysis (EDA) to explore any relationships between $\log(\mathbf{ai})$ and other factors. This establishes our expectations about relationships among the variables of interests. No confidence intervals or statistical tests are carried out in this exploratory chapter, these are included in Chapter 5 in our full regression models.

4.1 Counts of movers, stayers and intact families

In the 2001 Census, there were 3.7 million usual residents in New Zealand. When we look at the weighted data from the SoFIE, we find that there are 136,000 people in the New Zealand population, represented by 770 SoFIE OSMs, who are 'movers'. That is, they moved across TAs in wave 3, but not in any other wave. Conversely, 3,123,000 people, represented by 19,000 SoFIE OSMs, are 'stayers'. These people stayed in the same TA for all four waves. Note, this does not take into account movement between low level geographic areas, or other patterns of movement between TAs as indicated at the HED (these excluded scenarios are shown in Figure 3.1) (page 28).

As there are no family longitudinal weights provided in the micro dataset, here we present the unweighted counts of intact families in SoFIE samples to give an outline of the size of our group of interest and the control group. According to our definition of intact families, 5,890 have been identified in the SoFIE samples. As shown in Table 4.1 (page 46), 190 of these are mover intact families containing 400 OSMs (in which 75.0% of them

who are adults who aged 15 years and over); and 5,700 are stayer intact families, with 11,670 stayers OSMs (78.2% of them are adults).

Table 4.1: Unweighted counts of OSM movers and stayers in intact families

Migration_status	Total intact families	OSMs in intact families	Adult OSMs in intact families
Movers	190	400	300
Stayers	5700	11670	9130

4.2 Who are movers?

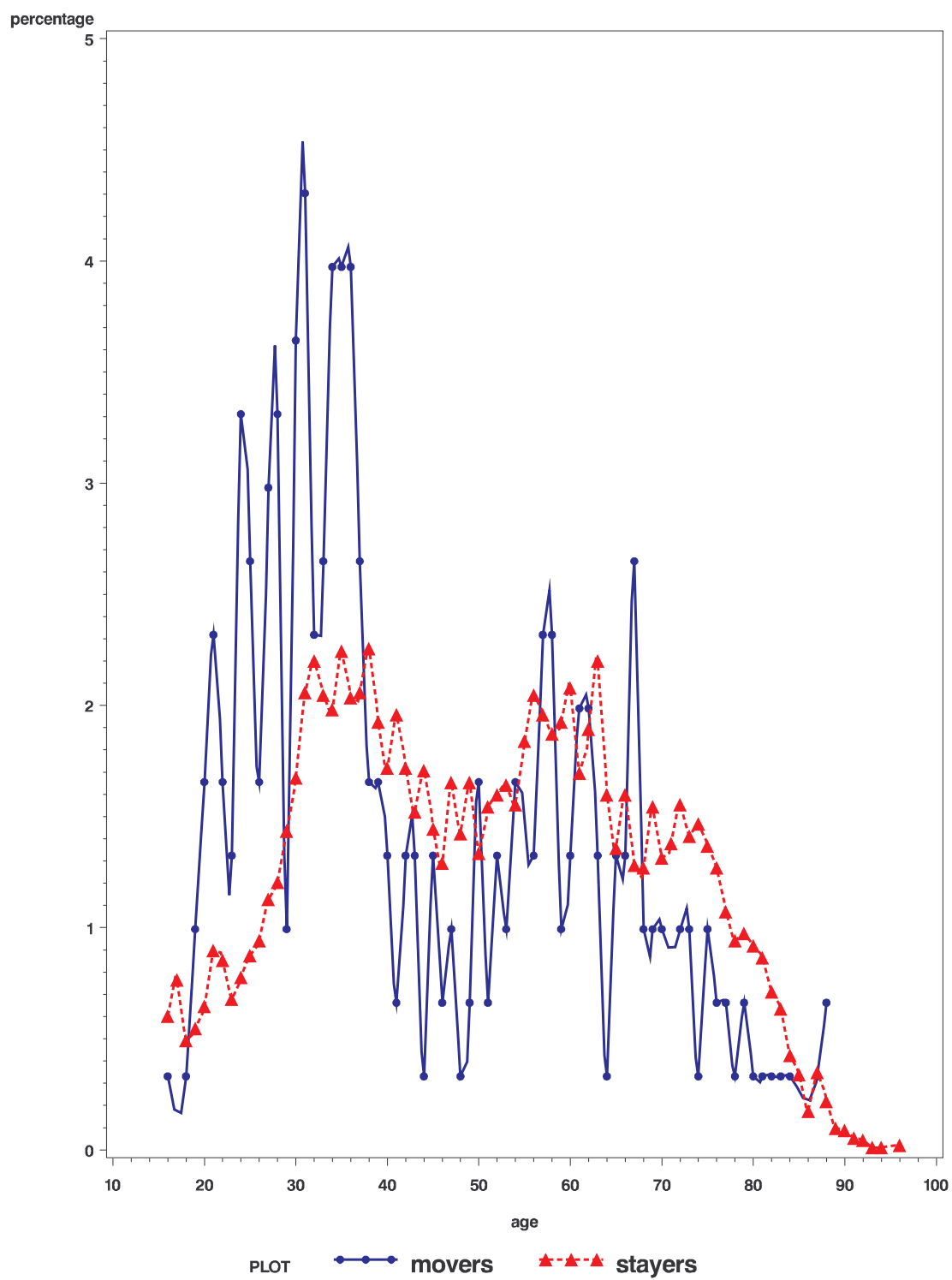
Before studying internal migration and its associated changes in $\log(\mathbf{ai})$, we would like to examine who those movers are and compare the individual characteristics of adult movers with adult stayers.

- Demographic characters: Age and Sex
Age and sex are basic demographic characteristics of individual respondents. Figure 4.1 (page 47) displays the difference in age* distribution in wave 2, between individual adults in the mover and stayer groups.

From Figure 4.1 (page 47) , we can see that both the age intervals 20 to 35 years, and 45 to 63 years have higher proportion in the mover group than in the stayer group; and the proportions of OSMs aged both 36 to 45 years, and 64 plus years, are higher in the stayer group than in the mover group. For the age group of 45-63, there are more likely to be families with children who have left home, giving the parents more freedom of moving. In contrast, for the age group 36-45, it could include many families with growing children, so the parents could be more likely to choose to stay. From Figure 4.1 we might

*Ages are **AgeAtHED** at wave 1, in the sample responding at wave 2

Figure 4.1: Percentage of adult movers and stayers by age, wave 2



expect that young adults move a lot, families with growing children do not, families with children who have left home move more frequently, and older people move less.

When looking at sex, there is not much difference between movers and stayers in the SoFIE sample. The proportion of males in both movers and stayers is slightly less than half of total number of movers and total number of stayers in the SoFIE sample, 44% for movers and 46% for stayers. The proportions of females in both movers and stayers are slightly more than a half of the total number of movers and the total number of stayers in the SoFIE sample, 56% for movers and 54% for stayers. It indicates to us that internal migration doesn't depend strongly on sex.

- Family type

From Table 4.2 (page 49) we can see that the spread of movers and stayers across family types is very similar. Most belong to these three types: couple only, couple with child(ren), and not in a family nucleus. In the mover group, only 3.2% are in a 'one parent with child(ren) family'. This is lower than the proportion in the stayers group, which is 6.1%. OSMs 'not in a family nucleus' are people who are living alone, or living with other people who are not members of their family, such as flatmates. Although we found that there is not a big difference in the distribution by family type between movers and stayers, it is very clear that one parent families more often stay rather than move.

- Education level

Table 4.3 (page 49) presents that the proportion of movers having a 'university qualification' is 6% higher than the proportion for stayers; the proportion of movers having 'no qualification' is 7% less than the proportion of stayers; and the proportions having a 'school or vocational qualification' is similar in both movers and stayers. This indicates that before internal migration, movers are more highly qualified than stayers.

Table 4.2: Adult OSMs in intact families by family types, wave 2

Family type	Movers (%)	Stayers (%)
Couple only	130 (41.9%)	3580 (39.2%)
Couple with child(ren)	100 (32.3%)	2960 (32.4%)
One parent with child(ren)	10 (3.2%)	560 (6.1%)
Not in a family nucleus	70 (22.6%)	2040 (22.3%)
Total	310 (100%)	9140 (100%)

Table 4.3: Highest qualification by migration status, wave 2

Qualification	Movers (%)	Stayers (%)
No school qual	60 (20.0%)	2510 (27.5%)
School or vocational qual	180 (60.0%)	5360 (58.6%)
University qual	60 (20.0%)	1270 (13.9%)
Total	300 (100%)	9140 (100%)

From the analysis above, we are able to draw a rough picture of movers. People in this group are more likely to be aged between 20 to 35, having a university degree, and not living in a one parent with child(ren) family.

4.3 Primary predictors of family annual income for intact families

To study changes in the financial situation of intact families, $\log(\text{ai})$ is our response variable, and time and migration status are our primary explanatory variables of interest. In this section we visualise the patterns between the response variable and these two explanatory variables. It is important to consider other explanatory variables, however, and we consider these in the next section.

As discussed in section 3.3 (page 36), family annual incomes have been transformed to $\log(\text{ai})$ for 190 mover and 5,700 stayer intact families. To visualise their relationships, firstly we just use simple random sampling to select 100 intact families and their corresponding $\log(\text{ai})$ changes from movers and stayers respectively, shown in Figure 4.2 (page 51). The graph suggests that more intact families in the mover group experienced fluctuations in $\log(\text{ai})$ than did intact families in the stayer group.

To get a clearer picture of $\log(\text{ai})$ changes between waves for intact families of movers and stayers, we calculate the mean values of $\log(\text{ai})$ for each wave, for these two groups separately. These are shown in Figure 4.3 (page 52), where the dashed red line represents the mean value change of $\log(\text{ai})$ of stayer intact families, and the solid blue line represents the change for mover intact families. Compared with a nearly constant rate of increase in $\log(\text{ai})$ in stayer intact families, mover intact families have a lower rate of increase between waves 2 and 3, which is their movement period, but a relatively higher increase between waves 1 and 2 and waves 3 and 4, which are the periods before and after their movement. The starting point for mover intact families is lower than stayer intact families, but at the end of wave 4 this difference has decreased. So, in general, within the first four waves, mover intact families have a higher rate of increase in $\log(\text{ai})$ than that of stayer intact families.

Figure 4.2: Family $\log(\mathbf{ai})$ changes for mover and stayer intact families
(Simple Random Sample of 100 families)

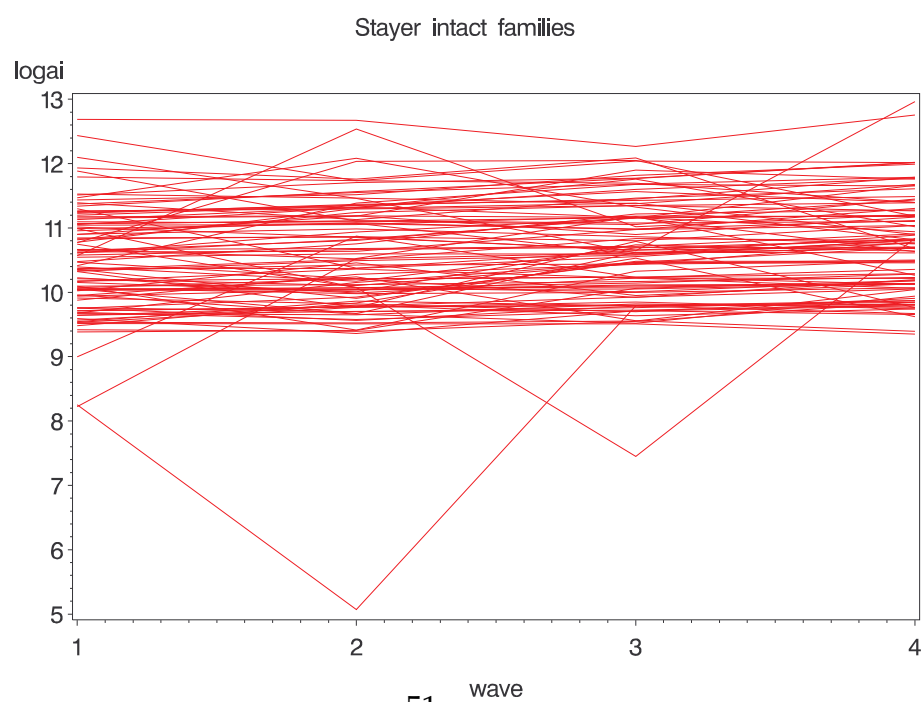
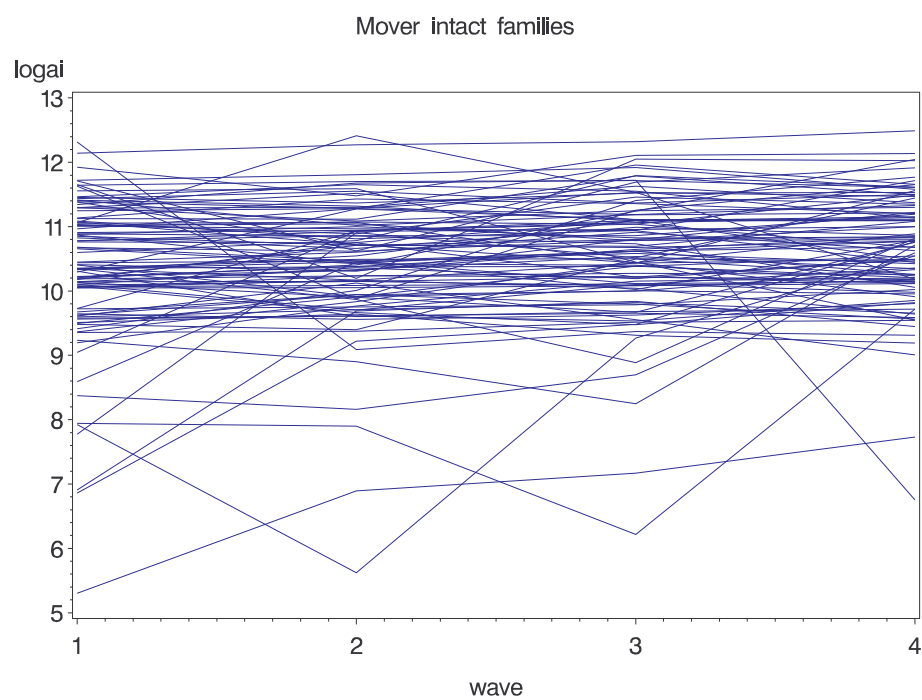
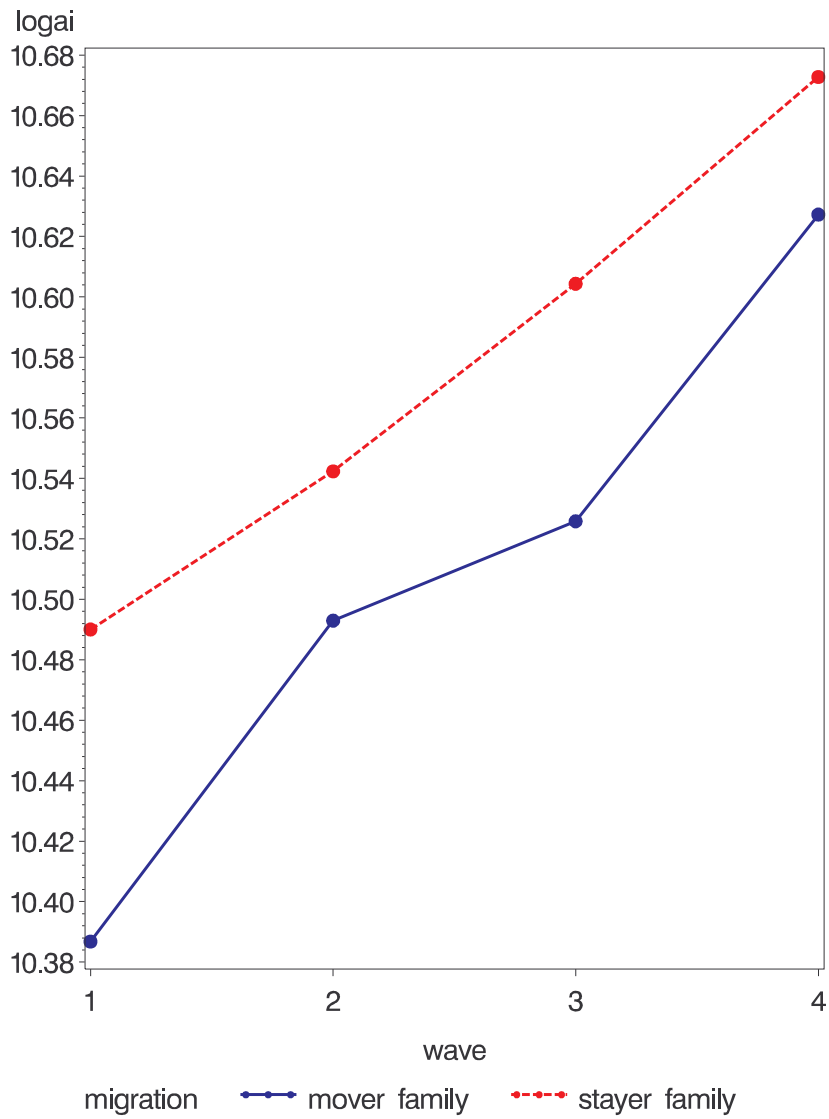


Figure 4.3: Mean of family $\log(\mathbf{ai})$



4.4 Secondary predictors of family annual income for intact families

Other predictors such as family type, average age of adult family members, the number of dependent children in the family, the highest qualification held by an adult member of the family, and family dwelling tenure status, may also play an important role in inferring the changes of $\log(\text{ai})$ over time. In this section, we study how these secondary predictors associate with $\log(\text{ai})$ as this helps to set up the model for analysis.

We analyse each of the predictors mentioned above separately. Firstly we show the distribution of $\log(\text{ai})$ by each category, using a box plot[†]. Then, with the categories baselined on wave 2, we compare the changes of the median value of $\log(\text{ai})$ of wave 2 to wave 4, between mover intact families and stayer intact families of each category to show the relationship between this predictor and response variable. Finally, we identify interactions between secondary predictors and primary predictors.

(1) Family type for an intact family (**familytype**)

Figure 4.4 shows that $\log(\text{ai})$ is similar across family types at wave 2. Couple with child(ren) families have the highest median $\log(\text{ai})$.

Figure 4.5 (page 55) shows that $\log(\text{ai})$ increases from wave 2 to wave 4 for 'Couple Only', 'Couple with Children' and 'Not in a family nucleus' intact mover families. But it decreases for 'One parent with children families'. For stayer intact families, all of the family type categories experience an increase in $\log(\text{ai})$ over time. Therefore, this suggests that changes in $\log(\text{ai})$ are affected by family type. From the graph of 'One parent with children', and 'Not in a family nucleus' families we can see that changes of $\log(\text{ai})$ also depend on the interactions between **casewave** and migration status. We therefore conclude that the three-factor interaction of **migration**, **casewave** and **familytype** should be considered in a full regression model.

[†]Note: due to confidentiality reasons, the points outside the upper and lower fences have not been displayed in the graph.

Figure 4.4: Family **log(ai)** by family type wave 2, where C/O: couple only; C/C: couple with child(ren); OP/C: one parent with child(ren); N/F: not in family nucleus

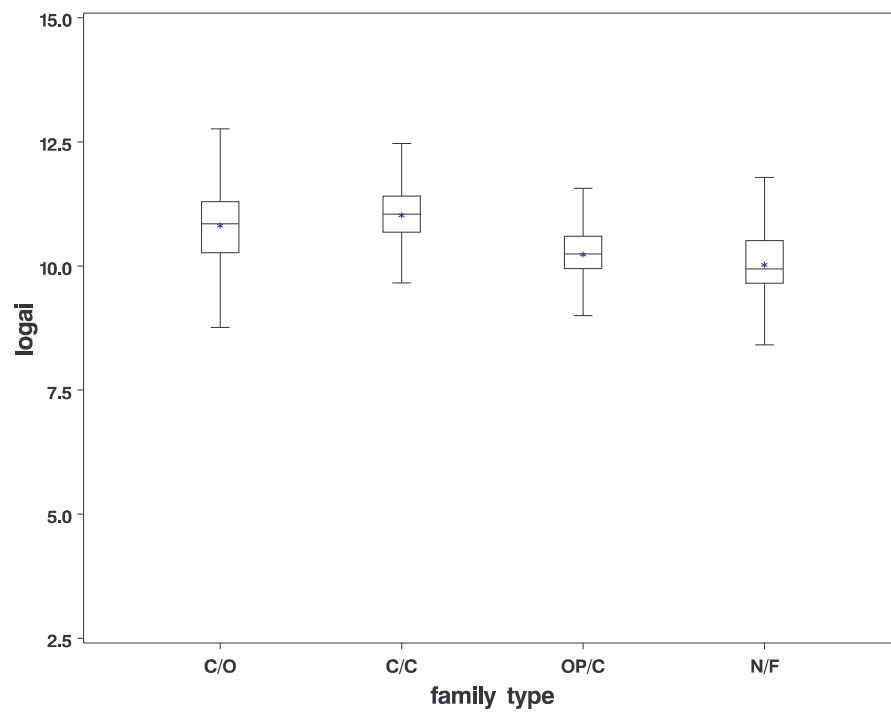
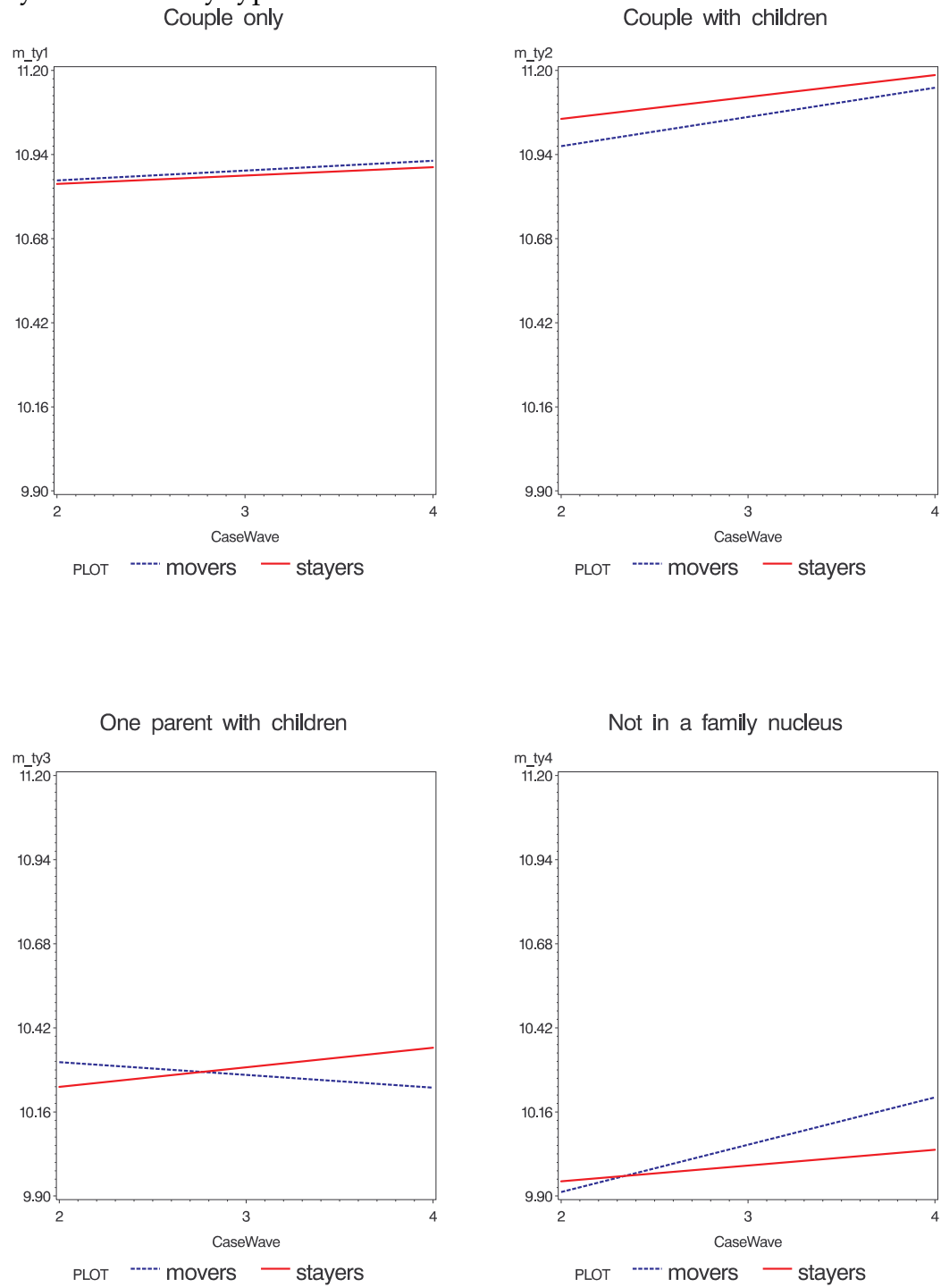


Figure 4.5: Changes in the median value of family **log(ai)** from waves 2 to 4, by wave 2 family type



(2) The average age of adult members in an intact family (**agecat** and **agebase**)

From Figure 4.6 (page 57) the median and mean values are in almost the same point for each age group. This shows that **log(ai)** is evenly spread within each group in **agecat**. Families where the average age of the adult members is within the two age groups of 26 to 45 years, and 46 to 65 years, have higher **log(ai)** than do those in both the youngest age group of 15 - 25 years, and the oldest age group of 66 years and over. We have produced Figure 4.7 (page 58) to show the non-monotonic relationship between **agebase** and **log(ai)**. From this graph we can see again that **log(ai)** is higher in wave 4 compared with wave 2. **log(ai)** is highest at age 39 for both waves, and is lower for both younger and older people.

Figure 4.8 (page 59) shows that the youngest age group shows an increase in **log(ai)** over time for both mover and stayer families. This is to be expected as new entrants to the work force gain experience and are promoted. The increase is greater for mover families than for stayer families. On the other hand, in older age groups, such as '46 to 65 years' and '66 and over', mover intact families decrease in **log(ai)** between waves 2 and 4. This is compared to an increase in **log(ai)** for stayer intact families of the same age. From the four graphs presented in Figure 4.8 (page 59) , we can see that the changes in **log(ai)** depend on the average age of adult members in an intact family and also on the relationship between migration status, case wave and the average age of adult family members, so the three-factor interaction should be considered when setting up the model. To make sure the model reflects the data, we use the continuous data **agebase** in our model.

Figure 4.6: Family $\log(\mathbf{ai})$ by **agecat**, wave 2

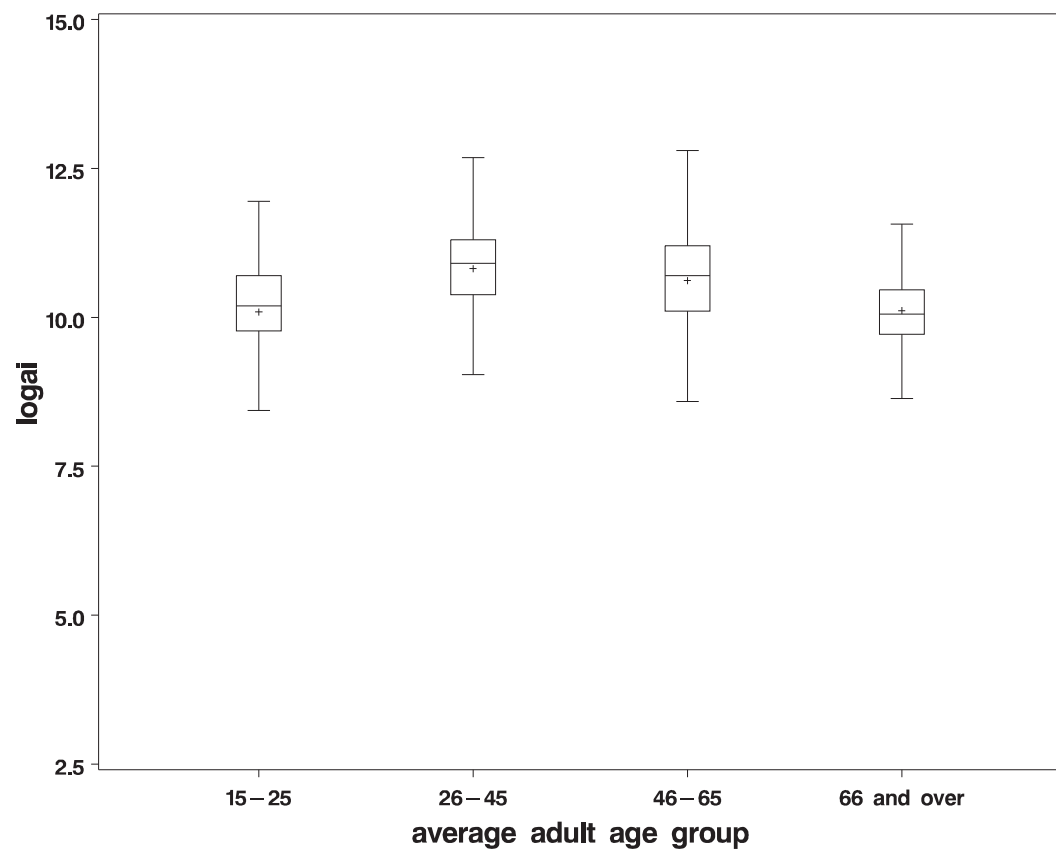


Figure 4.7: Family $\log(\mathbf{ai})$ by $\mathbf{agebase}$ in wave 2 and wave 4

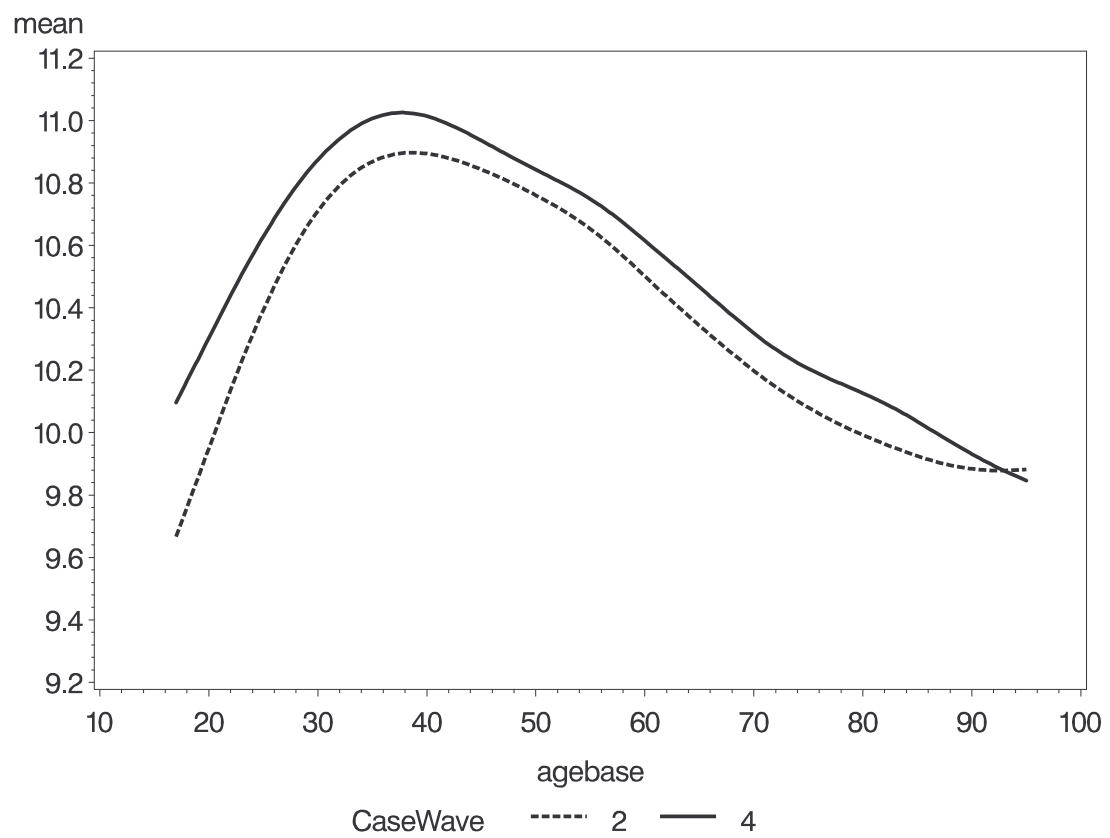
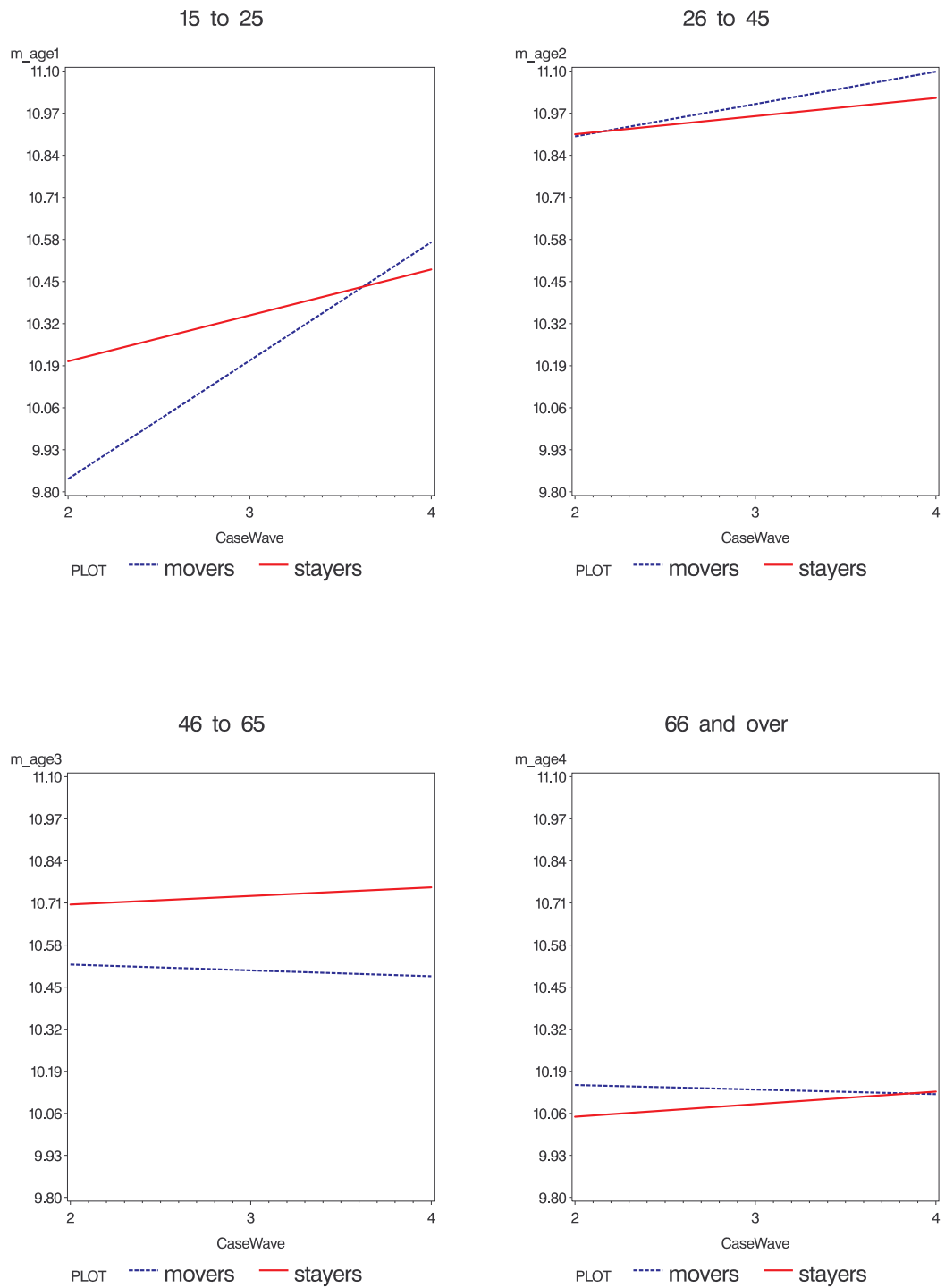


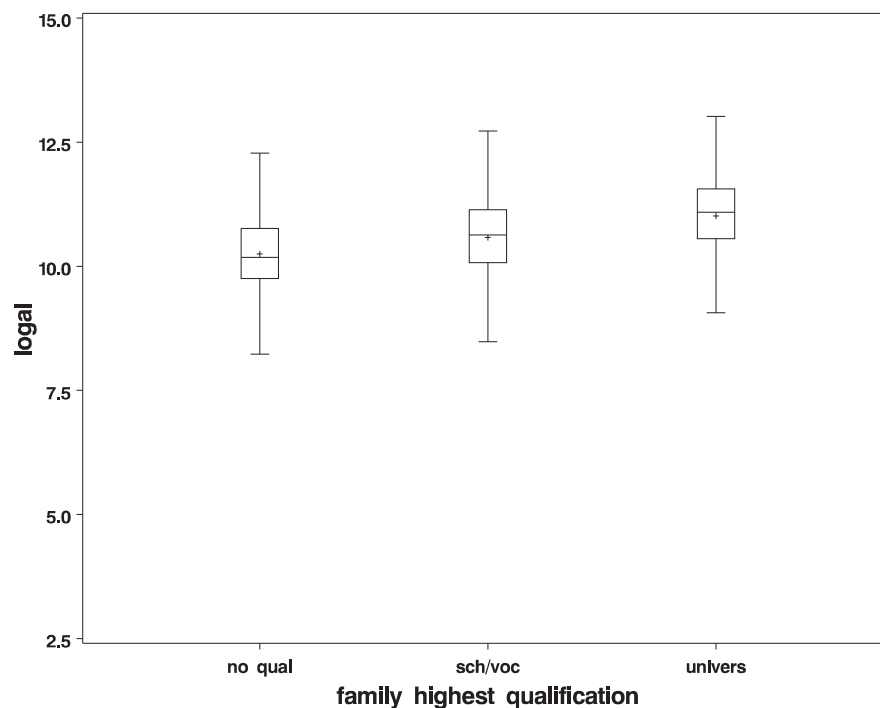
Figure 4.8: Changes in the median value of family $\log(\mathbf{ai})$ from wave 2 to 4, by wave 2 age groups (\mathbf{agecat}) of the average age of adults in the family



(3) Highest qualification of adults for an intact family (**highqual**)

The relationship between the highest level of education held by all adult family members and **log(ai)** is shown in Figure 4.9. This boxplot shows that **log(ai)** increases with increases in the highest qualification held by family members. The distribution of **log(ai)** is evenly spread in each level of education.

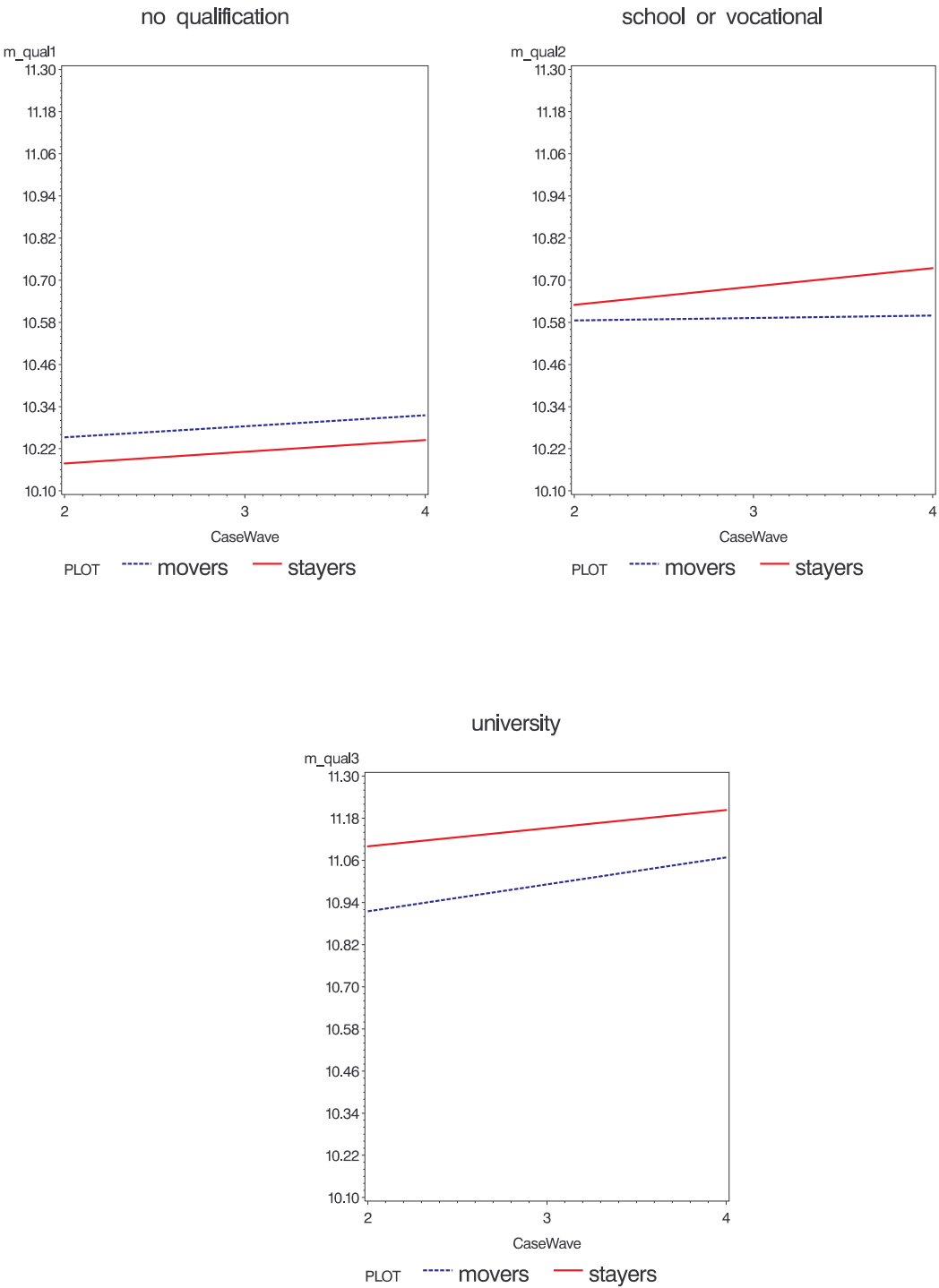
Figure 4.9: Family **log(ai)** by highest qualification of adult family members (**highqual**) wave 2, where no qual: no qualification; sch/voc: school or vocational qualification; univers: university qualification



In Figure 4.10 (page 61), we can see that both mover and stayer families with university qualifications have a significant increase in **log(ai)** over time when compared to families with lower qualifications. Only mover families with 'no qualification' have a higher **log(ai)** than stayer families. In all of the other categories, mover families have a lower **log(ai)** than stayer families. This suggests that the three-factor interaction between migration status, case wave, and highest qualification should be included in

the model.

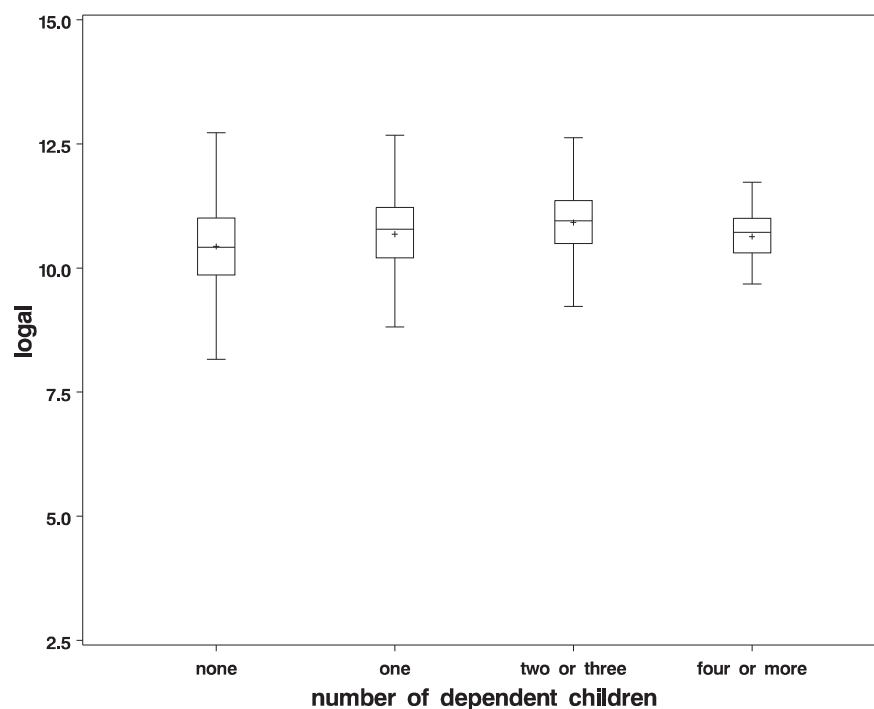
Figure 4.10: Changes in median value of family **log(ai)** from wave 2 to 4, by wave 2 highest qualification of adults (**highqual**)



(4) Number of dependent children for an intact family (**depc** and **depccat**)

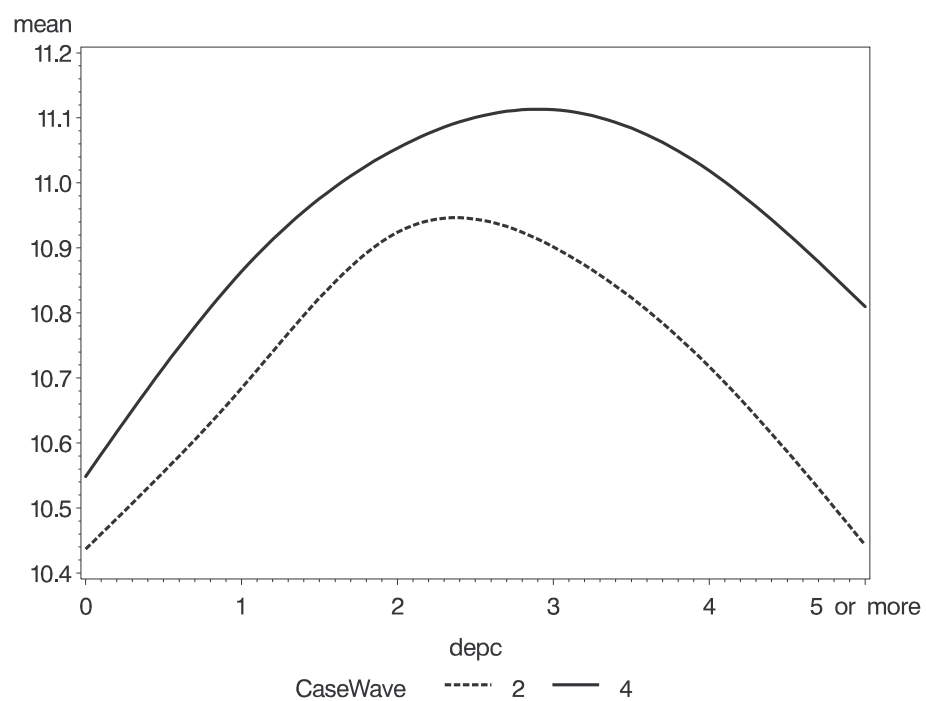
Figure 4.11 (page 62) shows that intact families with either no dependent children or four or more dependent children, have a lower **log(ai)** than families with between one and three dependent children. Similar to the age variable, the number of dependent children also shows a non-linear relationship with **log(ai)**, see Figure 4.12 (page 63).

Figure 4.11: Family **log(ai)** by number of dependent children (**depccat**), wave 2



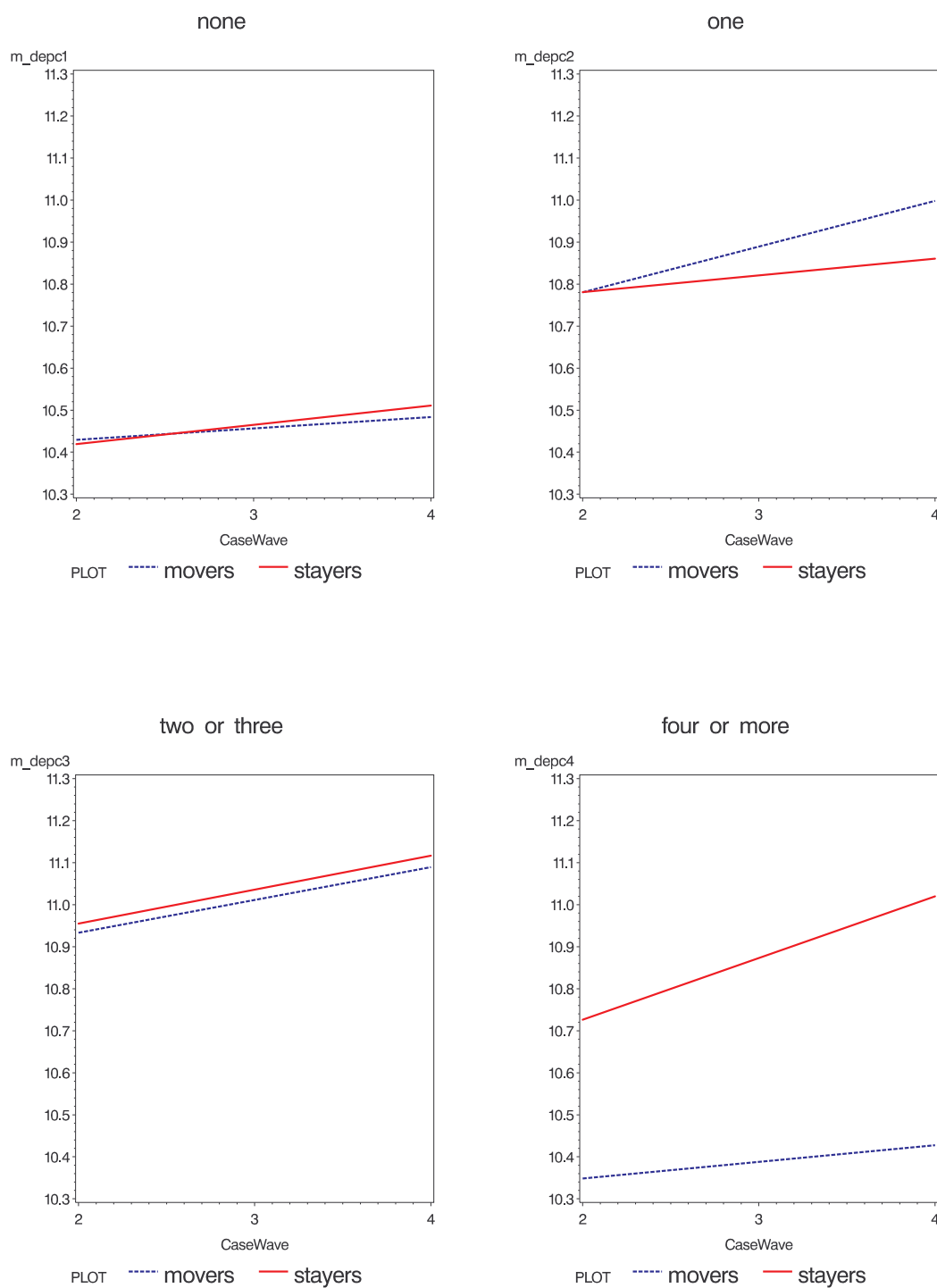
In Figure 4.13 (page 65), the increase in **log(ai)** from wave 2 to wave 4 is almost the same for movers and stayers in both families with no dependent children, and those with two or three dependent children. However, mover and stayer intact families experienced very different rates of **log(ai)** increase, when families had one dependent child, or four or more dependent children. This indicates that the impact of the number of dependent children on **log(ai)** depends on migration status and case wave. To see

Figure 4.12: Family $\log(\mathbf{ai})$ by number of dependent children (**depc**) for wave 2 and 4



if these are significant to **log(ai)** changes, we have added the three factor interaction of case wave and the number of dependent children and migration status into the model.

Figure 4.13: Changes in the median value of family $\log(\mathbf{ai})$ from waves 2 to 4, by wave 2 number of dependent children (**depccat**)



(5) Dwelling tenure status for intact families (**tenure**)

Table 4.4 and Table 4.5 (page 67) show that 52.6% of intact mover families owned their residential property in wave 2. This is slightly more than the 47.4% who were renting residential property. The proportions are very different for intact stayer families, with 71.9% owning their residential property, and only 28.1% renting. This indicates that fewer mover intact families own their own home before their move, than do stayer families during the same period. Although the proportion of both mover and stayer intact families who own their own home increases at wave 4, there are still fewer mover families who own their own dwelling. The difference in dwelling ownership between the two groups at this point is 20%. When the total number of families who own their home is compared with the number renting, we find no evidence that internal migration impacts dramatically on dwelling tenure. However, on closer inspection, it is evident that the transition of dwelling tenure is quite different for the two groups. There are 20% of mover intact families went from owning to renting in wave 4. This is a very high proportion, when compared with the 3.6% of stayer intact families who made the same transition. On the other hand, a higher proportion of mover intact families also changed from renting to owning than did stayer intact families (33.3% for movers and 27.5% for stayers). From this comparison we can say that changes in home ownership for both renting to owning, and owning to renting, are higher in movers intact families than stayers intact families over time.

To demonstrate the relationship between dwelling tenure and family **log(ai)** we have produced a box plot, shown in Figure 4.14. This illustrates that in wave 2, families who own their residential property have higher **log(ai)** than families who do not own their residential property.

From Figure 4.15 (page 69) we can see that both mover and stayer families who own their residential property have a higher **log(ai)** than do families who rent. Therefore, dwelling tenure status is an important factor for **log(ai)**. Intact families who own their own dwelling have experienced a dramatic change in their **log(ai)**. However, each of the groups moves into a different direction over time. Mover intact families experience a decrease in **log(ai)**, while stayer intact families experience an increase. This shows an interaction between dwelling tenure, migration status, and case wave.

Table 4.4: Dwelling tenure transition for intact families of movers

Dwelling tenure wave 2	Dwelling tenure wave 4		
	Owned type Row Pct Column Pct	Rented type Row Pct Column Pct	Total
Owned type Row Pct Column Pct	80 80.0% 72.7%	20 20.0% 25.0%	100 100.0% 52.6%
Rented type Row Pct Column Pct	30 33.3% 27.3%	60 66.7% 75.0%	90 100.0% 47.4%
Total Row Pct Column Pct	110 57.9% 100.0%	80 42.1% 100.0%	190

Table 4.5: Dwelling tenure transition for intact families of stayers

Dwelling tenure wave 2	Dwelling tenure wave 4		
	Owned type Row Pct Column Pct	Rented type Row Pct Column Pct	Total
Owned type Row Pct Column Pct	3770 96.4% 90.0%	140 3.6% 11.2%	3910 100.0% 71.9%
Rented type Row Pct Column Pct	420 27.5% 10.0%	1110 72.6% 88.8%	1530 100.0% 28.1%
Total Row Pct Column Pct	4190 77.0% 100.0%	1250 23.0% 100.0%	5440
Counts here excluded missing values			

Figure 4.14: Family $\log(\mathbf{ai})$ by dwelling tenure status wave 2 (**tenure**)

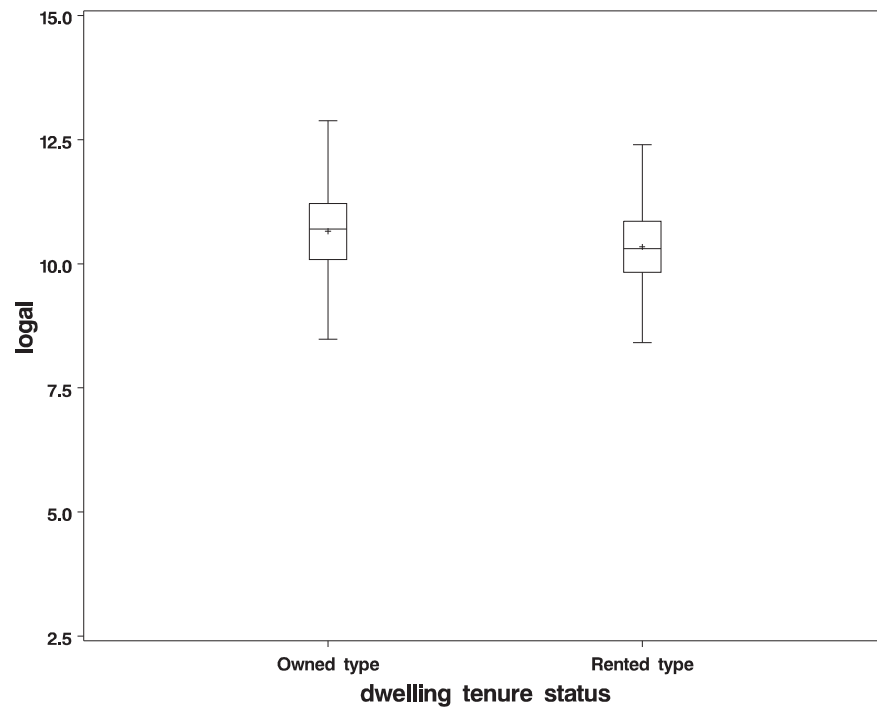
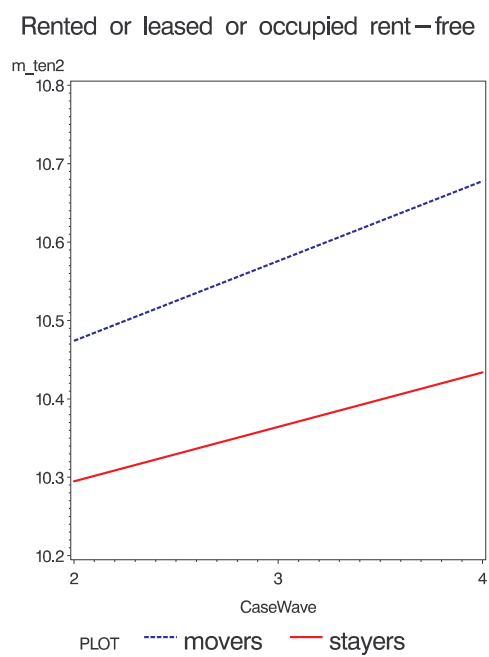
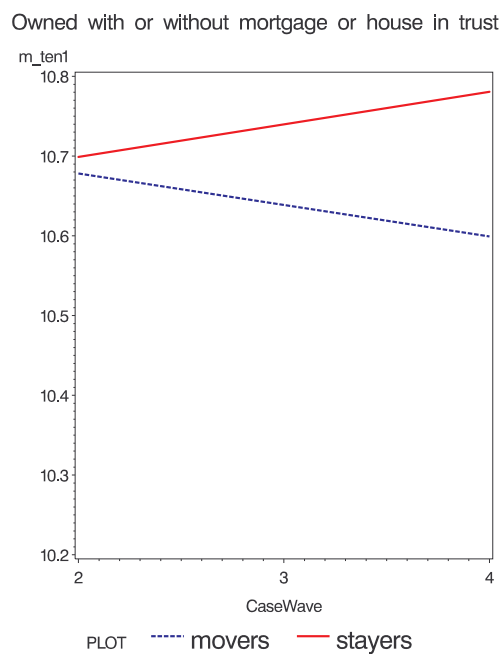


Figure 4.15: Changes in the median value of family **log(ai)** from waves 2 to 4, by wave 2 dwelling tenure status (**tenure**)



Chapter 5

Analysis and inference

In this chapter, we introduce the analysis method used in this thesis, and explain how to apply the method on the data we have. We then present our model selection and analyse the outputs produced by the best model.

5.1 Introduction analytical methods

This section introduces the statistical model used in the analysis for this thesis. Much of the content of this section is drawn from *Longitudinal Data Analysis with Discrete and Continuous Responses Course Notes* (2002), and Diggle et al. (2002).

5.1.1 General Linear Model

The common method of analysing cross-sectional continuous data is the General Linear Model which has a matrix form as below,

$$\underline{Y} = \mathbf{X}\underline{\beta} + \underline{\varepsilon} \quad (5.1)$$

In Equation (5.1), \underline{Y} is the vector of observed responses as below with independent component $Y_i \sim N(\mu_i, \sigma^2)$, where μ_i is the mean, $\mu_i = \underline{x}_i^T \underline{\beta}$ and σ^2 is constant variance.

$$\underline{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_i \\ \vdots \\ Y_n \end{pmatrix}$$

\mathbf{X} is the design matrix of predictors as below,

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ \vdots & & & \\ x_{i1} & x_{i2} & \cdots & x_{ip} \\ \vdots & & & \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

We define the vector

$$\underline{x}_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{pmatrix}$$

and \underline{x}_i^T represents the i^{th} row of the design matrix \mathbf{X} . $\underline{\beta}$ is the vector of regression parameters as below,

$$\underline{\beta} = \begin{pmatrix} \beta_1 \\ \beta_i \\ \vdots \\ \beta_p \end{pmatrix}$$

$\underline{\varepsilon}$ is the vector of random errors, $\underline{\varepsilon} \sim N(0, \sigma^2 I)$. Here we assume the error terms are independent and normally distributed with zero mean and common variance. Therefore the expected value of Y_i is as below,

$$\mu_i = E[Y_i] = \underline{x}_i^T \underline{\beta} = \begin{pmatrix} x_{i1} & x_{i2} & \cdots & x_{ip} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_i \\ \vdots \\ \beta_p \end{pmatrix} = x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{ip}\beta_p \quad (5.2)$$

Can we use the General Linear Model as described above to estimate longitudinal data? The answer is no. It is not appropriate and the inference is invalid. In a longitudinal data set, the response variables are repeated measurements for the same subjects over time. It is inappropriate to assume that the responses of the same subject are independent, as the measurements within the same subject tend to be correlated. The correlation may depend on time. For example, if responses relating to the same subject are measured at short intervals, they will be highly correlated. If the

responses are measured less frequently, they will be less correlated. These patterns of correlation and variation may combine to produce a complicated covariance structure. This means that special methods should be used when analysing longitudinal data. Here we introduce the General Linear Mixed Model.

5.1.2 General Linear Mixed Model

The General Linear Mixed Model is an extension of the General Linear Model. It allows random effect parameters to be added and also handles a more flexible specification of the covariance matrix of random errors, see the Equation (5.3).

$$\underline{Y} = \mathbf{X}\underline{\beta} + \mathbf{Z}\underline{\gamma} + \underline{\varepsilon} \quad (5.3)$$

In this model, \underline{Y} is the vector of observed responses with repeated measurements.

\mathbf{X} is the design matrix of fixed-effects.

$\underline{\beta}$ is the vector of fixed-effect parameters. It represents the parameters for fixed effects and reflects how the subject-specific regression coefficients relate to known covariates. $\underline{\beta}$ are assumed to be the same for all subjects.

\mathbf{Z} is the design matrix of random-effects.

$\underline{\gamma}$ is the vector of random-effect parameters. It represents the parameter for random effects. It contains separate components for each subject, including an individual effect (random intercept) as well as possible individual specific modification of the relationship between the outcome and certain predictors.

As stated by Diggle et al. (2002), $\underline{\varepsilon}$ is a vector of random errors which are not necessarily independent. For any subject i , we could divide $\underline{\varepsilon}_i$ into $\underline{\varepsilon}_i = \underline{\varepsilon}_{(1)i} + \underline{\varepsilon}_{(2)i}$, where $\varepsilon_{(1)i}$ is the measurement error reflecting the variation added by the measurement process and $\varepsilon_{(2)i}$ is the error associated with the serial correlation in which closer intervals are more correlated than intervals apart.

This model assumes that the random effects and error terms are normally distributed with means of 0; random effects and error terms are independent of each other; and that the relationship between the response variable and predictor variables is linear.

For subject i , Equation (5.3) is

$$\underline{Y}_i = \mathbf{X}_i \underline{\beta} + \mathbf{Z}_i \underline{\gamma}_i + \underline{\varepsilon}_i, \quad i = 1, \dots, n \quad (5.4)$$

Equation (5.3) also can be written as below:

$$\underline{Y} = \begin{bmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_i \\ \vdots \\ \mathbf{X}_n \end{bmatrix} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \mathbf{Z}_1 & & & \\ & \ddots & & \\ & & \mathbf{Z}_i & \\ & & & \ddots \\ & & & & \mathbf{Z}_n \end{bmatrix} \begin{bmatrix} \underline{\gamma}_1 \\ \vdots \\ \underline{\gamma}_i \\ \vdots \\ \underline{\gamma}_n \end{bmatrix} + \begin{bmatrix} \underline{\varepsilon}_1 \\ \vdots \\ \underline{\varepsilon}_i \\ \vdots \\ \underline{\varepsilon}_n \end{bmatrix} \quad (5.5)$$

In Equation (5.5), $\mathbf{X}_i, i = 1 \dots n$ is a $m * p$ design matrix for the fixed effects for subject i , m is the total number of observations of subject i , and p is the total number of fixed effects in the model.

$\underline{\beta}$ is the vector of regression parameters with p elements.

\mathbf{Z}_i is a $m * k$ design matrix of random effects for subject i , where m is total number of observations and k is total number of random effects.

$\underline{\gamma}_i$ is the vector of random effect parameters with k elements.

The covariance structure of the random effects is denoted as matrix \mathbf{G} and the covariance structure of random errors is denoted matrix as \mathbf{R} .

$$\text{Var}(\varepsilon) = \mathbf{R}$$

$$\text{Var}(\gamma) = \mathbf{G}$$

To calculate the variance of \mathbf{Y} ,

$$\text{Var}(\mathbf{Y}) = \text{Var}(\mathbf{Z}\gamma + \varepsilon) = \text{Var}(\mathbf{Z}\gamma) + \text{Var}(\varepsilon) = \mathbf{ZGZ}^T + \mathbf{R} \quad (5.6)$$

The method of estimation of \mathbf{G} and \mathbf{R} is Restricted Maximum Likelihood (REML). Maximum Likelihood Estimation (MLE) and REML often give very similar results, the distinction between MLE and REML estimation is important only when p is relatively large*.

*Diggle et al. (2002), Section 4.5 contains the detailed information on this method of estimation

5.1.3 Covariance structure for \mathbf{R}

Matrix $\mathbf{R} = \text{Var}(\underline{\varepsilon})$ has a block-diagonal covariance structure where the block corresponds to the covariance structure for each subject. The observations within each block can take on a variety of covariance structures while the observations outside of the block are assumed to be independent. To get more accurate estimates, the selection of an appropriate covariance structure of the blocks of \mathbf{R} is important. There are five structure options. The independent structure is the simplest option. This assumes that the within-subject correlations are zero. See the example matrix below when $m = 3$:

$$\begin{bmatrix} \sigma^2 & 0 & 0 \\ & \sigma^2 & 0 \\ & & \sigma^2 \end{bmatrix}$$

(Note that the blank indicates a symmetrical component, which is also applied in the following matrices.) This structure is suitable for measuring changes between very long time intervals, where the within-subject correlation can be treated as zero.

Another simple correlation structure is the compound symmetry structure, see the matrix below,

$$\begin{bmatrix} 1 & \rho & \rho \\ & 1 & \rho \\ & & 1 \end{bmatrix}$$

In this structure, within-subject correlation is assumed to be constant, regardless of the distance between the measurement time points. This structure is suitable when the correlation doesn't depend on time lag.

Unstructured covariance structure is the most complex structure, in which the measurements relating to each pair of time periods have their own unique correlations, see the matrix below,

$$\begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ & \sigma_2^2 & \sigma_{23} \\ & & \sigma_3^2 \end{bmatrix}$$

An unstructured covariance structure requires estimation of a large number of variance components, and it will lead to a reduction in both the efficiency and power of tests for the model parameters.

The first-order autoregressive AR (1) structure makes the correlation between observations a function of the number of time points apart, see the matrix below,

$$\begin{bmatrix} 1 & \rho & \rho^2 \\ & 1 & \rho \\ & & 1 \end{bmatrix}$$

This structure assumes the distance of two conjacent time points is equal, with correlation of a constant amount ρ .

There are also other covariance structures that allow for unequal spacing, such as a spatial power structure. In this structure the power of each of ρ is the absolute value of difference between two conjacent time points, see the matrix below,

$$\begin{bmatrix} 1 & \rho^{|t_1-t_2|} & \rho^{|t_1-t_3|} \\ & 1 & \rho^{|t_2-t_3|} \\ & & 1 \end{bmatrix}$$

This structure shows that the correlations decrease while the length of time increases.

5.2 Applying the General Linear Mixed model in our case

To apply the General Linear Mixed Model, we need to study the source of variations in the family **log(ai)** to choose an appropriate covariance structure.

5.2.1 Source of random variation

The covariance structure is decided by the source of random variation. Sources of the variation are random effects, serial correlation and measurement error.

Random effects come from consistent differences between individuals which can not be attributed to differences in measured covariance.

Serial correlation is the time-varying stochastic process operating within the subject. For example, the changes in the observed measurement of family **log(ai)** may be due to the correlation between pairs of measurements relating to the same subject, and those correlations tend to weaken as the time between observations increases.

Measurement error is the variation from the measurement process. This may be an error occurring in the data collection or transformation process.

5.2.2 Selecting covariance structure for our data

This thesis uses a variogram to quantify the three sources of variation; this is an exploratory exercise used to select an appropriate covariance structure. The steps used to produce a variogram are as follows:

Step one, calculate pairwise squared residual differences.

As we have i subjects and j measurements for each subject, we fit the general linear model of Equation (5.1)

$$y_{ij} = \underline{x}_{ij}^T \underline{\beta} + \varepsilon_{ij}, \quad i = 1, \dots, n, \quad j = 1, \dots, 4$$

This model assumes independent error terms $\varepsilon_i \sim N(0, \sigma_\varepsilon^2 I)$, *iid*, and that the response variable **log(ai)** is related to the identified predictors:

- **migration** (migration status),
- **casewave** (wave),
- **familytype** (type of family nucleus),
- **depc** (number of dependent children in the family) and its quadratic form as non-linear relationship has shown in Figure 4.12 (page 63),
- **highqual** (highest qualification held in the family),

- **tenure** (dwelling tenure)
- **agebase** (average age of adult members in the family in wave 2) and its quadratic form as non-linear relationship has shown in Figure 4.7 (page 58).

So, for the General Linear Model, we use the saturated model including all of the predictors and interactions within those predictors regardless of any within-subject correlation, to get the best estimate of the coefficients from the predictors. Model is as follows:

$$\begin{aligned} &\text{migration} * \text{familytype} * \text{agebase} * \text{agebase}^2 * \\ &\text{depc} * \text{depc}^2 * \text{highqual} * \text{tenure} * \text{casewave} \end{aligned} \quad (5.7)$$

After fitting this model we get the estimates for $\hat{\beta}$, calculate \hat{Y}_{ij} , and the residuals $r_{ij} = Y_{ij} - \hat{Y}_{ij}$. Then we calculate pairwise squared residual differences V of all given residuals, Diggle et al. (2002) shows the calculation equation as below ,

$$V_{ij i' j'} = 0.5((r_{ij} - r_{i' j'})^2)$$

and also the related wave lags δ

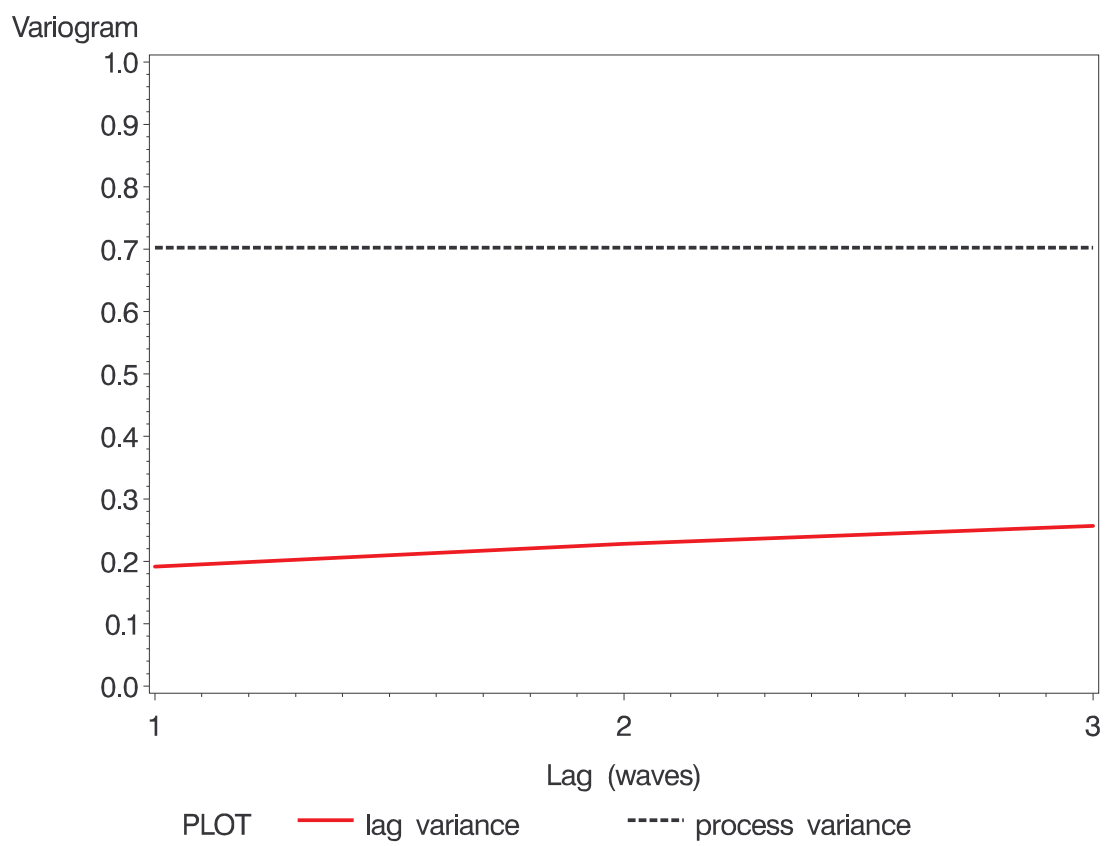
$$\delta_{ij i' j'} = |t_{ij} - t_{i' j'}|$$

Step two, calculate process variance. This is the mean value of all pairwise squared residual differences, except those that are within the same subjects (ie. exclude $i = i'$). This single value is shown as the dashed black line at the top of Figure 5.1 (page 78).

Step three, calculate the means of the squared residual differences within individuals at each wave lag $|t_j - t_{j'}|$. This is shown as the solid red line in the Figure (5.1) (page 78).

Here based on the theory from Diggle et al. (2002), we use Figure 5.1 to identify the sources of variation . The measurement error is shown as the vertical distance between zero and the starting point of the solid red line

Figure 5.1: Variogram of family $\log(\mathbf{ai})$



representing the means of squared residual differences. The serial correlation is represented by the slope of solid red line. In another words, it is the vertical distance between the start at lag point 1 and end at lag point 3 of this line. The existence of random effects are shown by the vertical distance between this line and the process variance line at the highest lag available. If no random effects were necessary in the model, the solid red line would rise to meet the process variance. Figure 5.1 (page 78) indicates that random effects are necessary.

The General Linear Mixed model with random coefficients is therefore used in the following analysis. To efficiently analyse the data, we used the simplest model, in which the intercept is the only random coefficient. In this model we assume each intact family has a random starting point of $\log(\mathbf{a}_i)$ in their first wave. The results from the variogram shows that the model with random intercept is plausible. However, for future research, it is worth exploring different models with different random effects.

5.2.3 Setting up the General Linear Mixed Model

In the model, we denote an intact family as i , where $i = 1, 2, \dots$. We denote wave as j where $j = 1, 2, 3, 4$. The General Linear Mixed Model with the intercept as A random effect becomes:

$$Y_{ij} = \underline{x}_{ij}^T \underline{\beta} + \gamma_i + \varepsilon_{ij} \quad (5.8)$$

Here the assumption is that given γ_i , Y_{ij} and Y_{ih} are independent for $j \neq h$. The model is equal to the model without any random effects that assumes the covariance structure for the error term is compound symmetry. See the proof as below. The model Equation (5.8) is presented as:

$$\underline{Y}_i = \mathbf{X}_i \underline{\beta} + \mathbf{1} \gamma_i + \underline{\varepsilon}_i$$

$$\text{Var}(\underline{Y}_i) = \text{Var}(\gamma_i \mathbf{1}) + \text{Var}(\underline{\varepsilon}_i) = \mathbf{1} \text{Var}(\gamma_i) \mathbf{1}^T + \text{Var}(\underline{\varepsilon}_i)$$

as $\underline{\varepsilon}_i \sim N(0, \sigma_\varepsilon^2 I)$ and $\gamma_i \sim N(0, \sigma_\gamma^2 I)$ and they are independent. Therefore,

$$\begin{aligned}
\text{Var}(\underline{Y}_i) &= \sigma_\gamma^2(\underline{1}\underline{1}^T) + \sigma_\varepsilon^2 I \\
&= \sigma_\gamma^2 \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} + \sigma_\varepsilon^2 \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \\
&= \begin{bmatrix} \sigma_\gamma^2 + \sigma_\varepsilon^2 & \sigma_\gamma^2 & \sigma_\gamma^2 & \sigma_\gamma^2 \\ \sigma_\gamma^2 & \sigma_\gamma^2 + \sigma_\varepsilon^2 & \sigma_\gamma^2 & \sigma_\gamma^2 \\ \sigma_\gamma^2 & \sigma_\gamma^2 & \sigma_\gamma^2 + \sigma_\varepsilon^2 & \sigma_\gamma^2 \\ \sigma_\gamma^2 & \sigma_\gamma^2 & \sigma_\gamma^2 & \sigma_\gamma^2 + \sigma_\varepsilon^2 \end{bmatrix}
\end{aligned}$$

Let

$$\rho = \frac{\sigma_\gamma^2}{\sigma_\gamma^2 + \sigma_\varepsilon^2}$$

Therefore,

$$\begin{aligned}
\text{Var}(\underline{Y}_i) &= (\sigma_\gamma^2 + \sigma_\varepsilon^2) \begin{bmatrix} 1 & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho \\ \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & 1 \end{bmatrix} \\
&= \sigma^2 \begin{bmatrix} 1 & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho \\ \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & 1 \end{bmatrix}
\end{aligned}$$

where

$$\sigma^2 = \sigma_\gamma^2 + \sigma_\varepsilon^2$$

Here is the model with compound symmetry as the covariance structure of $\underline{\varepsilon}$ and without random effects,

$$Y_{ij} = \underline{x}_{ij}^T \beta + \varepsilon_{ij}$$

or,

$$\underline{Y}_i = \mathbf{X}_i \beta + \underline{\varepsilon}_i$$

as $\underline{\varepsilon}_i \sim N(0, \sigma^2 W)$, where

$$W = \begin{bmatrix} 1 & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho \\ \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & 1 \end{bmatrix}$$

therefore,

$$\text{Var}(\underline{Y}_i) = \text{Var}(\underline{\varepsilon}_i) = \sigma^2 W$$

$$= \begin{bmatrix} \sigma^2 & \rho\sigma^2 & \rho\sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 & \rho\sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \rho\sigma^2 & \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \rho\sigma^2 & \rho\sigma^2 & \sigma^2 \end{bmatrix}$$

$$= \sigma^2 \begin{bmatrix} 1 & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho \\ \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & 1 \end{bmatrix}$$

From the model above we can see that the model with the intercept as the only random effect, and the model with compound symmetry as the covariance structure of the random error term give the same diagonal block matrix. The model implies that within a family the serial correlation is a constant ρ and the correlation between families are 0. As was shown in Figure 5.1 (page 78), this illustrates that serial correlations are small and almost constant. In this thesis we tried to both simplify and accurate our model, therefore compound symmetry is the best option for us. The possibilities of using other covariance structures can be explored in the future study.

5.2.4 Using the PROC MIXED procedure from SAS

The PROC MIXED procedure from SAS is used in this thesis to do the longitudinal analysis and inference for the continuous dependent variable **log(ai)**. The contents in this section are referenced from The SAS/STAT(R) User's GuideSAS/STAT(R) 9.22 *User's Guide* (2010). The statements used

in our analysis is discussed below. Statements such as 'BY', 'CONTRAST', 'ESTIMATE' and so on, which are not used in this analysis, are not discussed here.

- **The syntax of the procedure used is given below :**

```
PROC MIXED DATA= ;
CLASS ;
MODEL / SOLUTION DDFM=KR;
RANDOM / SUBJECT= ;
LSMEANS /DIFF= ;
RUN;
```

In this procedure, PROC MIXED is to call the proc mixed procedure.

DATA= is to specify the dataset used in this procedure.

The CLASS statement specifies categorical variables in those predictors.

The MODEL statement specifies the model of $\log(\mathbf{a}_i) = \text{function of predictors}$, which is $\underline{x}_{ij}^T \underline{\beta}$ in Equation (5.8).

The RANDOM statement specifies the random effects in the model, here we treat the intercept as the random effect, which is γ_i in Equation (5.8).

The SOLUTION option requests estimates for all fixed effects in the model, together with the standard errors, t-statistics, and p-values. The DDFM=KR option requests the degrees of freedom calculations proposed by Kenward and Roger (1997).

The SUBJECT option identifies the family ID in the model.

The REPEATED statement is used to specify the covariance structure of random error \mathbf{R} , which is $Var(\underline{\varepsilon}_{ij})$ of $\underline{\varepsilon}_{ij}$ in Equation 5.8. Since we use the random intercept model, we do not need the REPEATED statement. All within-subject correlation is described by the random intercept.

- **LSMEANS statement**

The LSMEANS statement and related options is used in this thesis to calculate the difference of Least Square Means of **log(ai)** between different migration statuses over time. In *SAS/STAT(R) 9.22 User's Guide* (2010), it states that 'each LS-mean is computed as $\mathbf{L}\hat{\underline{\beta}}$, where \mathbf{L} is a contrast matrix associated with the least squares mean and $\hat{\underline{\beta}}$ is the estimate of the fixed-effects parameter vector'. The option **DIFF** specifies requested calculations on the difference of least squares means. Here is a simple example to describe how the LSMEANS works. Assume we have a model

$$E(Y) = \mu + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 \quad (5.9)$$

Where Y is **log(ai)** and it only relates to **migration** (x_1) and **casewave** (x_2) and their interaction $x_1 x_2$;

μ is the intercept;

β_1 is the main effect of **migration**, and x_1 is a dummy variable where 1 represents mover and 0 represents a stayer;

β_2 is the main effect of **casewave**, and x_2 is a dummy variable where 1 represents wave 4 and 0 represents wave 2;

β_{12} is the effect of the two-factor interaction of x_1 and x_2 .

Therefore the Least Square Means of Equation (5.9) for different combinations of these two factors are shown in Table 5.1,

Table 5.1: Migration status and wave effects table

Migration status	Wave	
	$x_2 = 1$	$x_2 = 0$
$x_1 = 0$	$E(Y_{sw4}) = \mu + \beta_2$	$E(Y_{sw2}) = \mu$
$x_1 = 1$	$E(Y_{mw4}) = \mu + \beta_1 + \beta_2 + \beta_{12}$	$E(Y_{mw2}) = \mu + \beta_1$

In Table 5.1,

- $E(Y_{sw4})$ is the mean of **log(ai)** for stayer families in wave 4;
- $E(Y_{sw2})$ is the mean of **log(ai)** for stayer families in wave 2;

- $E(Y_{mw4})$ is the mean of **log(ai)** for mover families in wave 4;
- $E(Y_{mw2})$ is the mean of **log(ai)** for mover families in wave 2;

To calculate the differences of Least Square Means of **log(ai)** between mover families and stayer families in wave 2 and wave 4 respectively we use 'DIFF=' in SAS to get the results below:

$$\hat{Y}_{mw2} - \hat{Y}_{sw2} = \hat{\beta}_1$$

which represents the difference of Least Squared Means of **log(ai)** between mover families and stayer families in wave 2.

$$\hat{Y}_{mw4} - \hat{Y}_{sw4} = \hat{\beta}_1 + \hat{\beta}_{12}$$

which represents the difference of Least Squared Means of **log(ai)** between mover families and stayer families in wave 4.

Since the final model we consider contains many three-factor interactions and continuous predictors, the difference of Least Squared Means of **log(ai)** between mover and stayer families has a complicated form. It can not be written as a simple expression of $\hat{\beta}_1$, $\hat{\beta}_2$ and $\hat{\beta}_{12}$. Therefore, it is easier to get the information using the LSMEANS statement in SAS. LSMEANS gives the difference between mover and stayer families averaging over all other predictors.

5.3 Model selection and analysis

In the initial model, the response variable is **log(ai)**, and predictors include the main effects of **migration**, **casewave**, **familytype**, **depc**, **depc²**, **highqual**, **tenure**, **agebase**, **agebase²**, and all of the two-factor interactions except the interaction between **depc** and **depc²** and **agebase** and **agebase²**, and the three-factor interactions between **migration**, **casewave** and other factors respectively. The SAS code and output of this initial model for waves 2 and 4 data is given in Appendix B.

Based on the concept of backward elimination selection, we did a manual

model selection to achieve the best model. As we are not using weighted data, the Type 3 test will be sufficient for the significance test. We set up the significance level at 5%, then items with P-value greater than 0.05 in the Type 3 test for the initial model are manually eliminated on an individual basis starting from the three-factor interactions. For example, the P-value of the three-factor interaction of **migration**, **familytype** and **casewave** is the highest value (among the three-factor interactions), so we eliminated this item first. We can then also eliminate the two-factor interactions of **familytype** and **migration**, and **familytype** and **casewave**, if they have a P-value higher than 0.05. This method was used to produce the final model shown in Appendix C for wave 2 and wave 4 data and Appendix D for wave 1, 2, 3 and 4 data.

Comparing the final model with the initial model for wave 2 and 4 data, we found that the final model has a smaller AIC value (21252.7) than the initial model (21504.3). This indicates that the final model provides a better fit for the data, which leads to more accurate and powerful inference. Here we summarise the Type 3 test of the final model in Table 5.2 (page 86) for wave 2 and 4 data and wave 1, 2, 3 and 4 data separately.

The summarised information in Table 5.2 (page 86) shows us that the relationships between key predictors and the response variable are similar in these two sets of data. The differences are the relationship between **log(ai)** and the two-factor interaction of **tenure * casewave** (only significant in wave 1-4 data), and **depc * casewave** and **depc² * casewave** (only significant in wave 2 and 4 data). The wave 1-4 data has more observations, which gives more opportunities for variation. We are interested in the immediate changes in **log(ai)** after the movement, not the impacts from confounding in 4 waves, so only wave 2 and wave 4 data is used in the following analysis.

Table 5.2: Type 3 Test of Fixed Effects, final model for wave 2 and 4 and for waves 1 to 4

Effect	P-value Wave 2 and 4 data	P-value Wave 1 to 4 data
depc	0.0003	<0.0001
depc ²	<0.0001	0.0003
agebase	0.0050	0.0002
agebase ²	0.0013	0.0001
migration	0.5555	0.2313
familytype	0.0023	<0.0001
highqual	0.0362	0.0040
tenure	<0.0001	<0.0001
casewave	<0.0001	<0.0001
depc * agebase	0.0031	0.0002
depc * agebase ²	0.0037	0.0002
agebase * migration	0.6159	0.3702
agebase ² * migration	0.4830	0.3603
depc * familytype	0.0003	0.0004
depc ² * familytype	<0.0001	<0.0001
agebase * familytype	0.0061	0.0252
agebase ² * familytype	0.0002	0.0010
depc * highqual	<0.0001	<0.0001
depc ² * highqual	<0.0001	<0.0001
agebase * highqual	<0.0001	<0.0001
agebase ² * highqual	<0.0001	<0.0001
familytype * tenure	0.0056	0.0494
depc * casewave	0.2468	
depc ² * casewave	0.0398	
agebase * casewave	<0.0001	<0.0001
agebase ² * casewave	<0.0001	<0.0001
migration * casewave	<0.0001	0.0002
tenure * casewave		0.0059
agebase * migration * casewave	<0.0001	0.0013
agebase ² * migration * casewave	0.0003	0.0084
blank indicates the item is not in the final model		

5.4 Estimation and inference

In this section we discuss the outputs from the final model for wave 2 and wave 4 data, and estimate the impact of the predictors related to our factors of interest, migration and wave, to make an inference on $\log(\mathbf{ai})$.

5.4.1 Estimates on main effects

- Key for variables
In Table 5.3 (page 89) we summarise estimates of the main effects given in Table C.3 (page 124), Appendix C, where depc , depc^2 , agebase and agebase^2 are continuous variables and the key for categorical variables is below,
 - migration
1=mover family,
2=stayer family;
 - familytype
1=not in a family nucleus,
2=one parent with child(ren),
3=couple only,
4=couple with child(ren);
 - highqual
1=no qualification,
2=school or vocational qualification,
3=university qualification;
 - casewave
four=wave 4
two=wave 2;
- Exponentiation
In Table 5.3 (page 89), let X be a dummy variable with value 1 and 0. Suppose the effect associated with X is β . Then

$$\begin{aligned}\hat{\beta} &= \log(E(\mathbf{ai})|X = 1) - \log(E(\mathbf{ai})|X = 0) \\ &= \log \frac{E(\mathbf{ai}|X = 1)}{E(\mathbf{ai}|X = 0)}\end{aligned}$$

where $E(ai|X = 1)$ is the family annual income given X on level 1, $E(ai|X = 0)$ is the family annual income given X on level 0, therefore, the exponentiation of $\hat{\beta}$ is,

$$e^{\hat{\beta}} = \frac{E(ai|X = 1)}{E(ai|X = 0)}$$

which indicates the ratio of the family annual income between level 1 and the reference level (level 0) of this main effect. For the continuous variable, the ratio indicates the changes happening in the family annual income when the variable increases by 1 unit. Then we use the exponentiation to calculate the Confidence Interval (C.I.) of the ratio,

$$C.I. = (e^{\hat{\beta}-1.96SE}, e^{\hat{\beta}+1.96SE})$$

where SE is the Standard Error of $\hat{\beta}$. This exponentiation method is also used in the following analysis of model outputs.

- Interpretation of Estimates

In general, given other effects are same, all families earned more income in wave 4 than did those families in wave 2, the ratio between these two waves is 1.4192. Annual incomes for mover families are 38.22% of that of stayer families. Families who own their residential properties got 1.1618 times of the annual income than did those families who rent their residential properties. If looking at the highest qualification of a family, the higher the qualification is the more the family annual incomes are. The annual incomes for families with school or vocational qualification or no qualification are 63.55% or 54.57% of that for families with university qualifications. Those results reflect our expectations from EDA.

However, the relationship between family type and **log(ai)** is shown differently in the output of the final model and in the EDA results. In EDA, a family of 'couple with child(ren)' family type with a higher income than other family types, but on the contrary, their incomes are the lowest in the output of the final model. Here we notice that each family type doesn't show a strong relationship with **log(ai)** based on their P-values in Table 5.3 (page 89). But from the Appendix C.3 (page 124) its interactions between average age of adult

Table 5.3: Summary of estimates of main effects in final model

Main Effect	Category	Estimates $\hat{\beta}$	Ratio to reference level $e^{\hat{\beta}}$	Confidence interval of ratio ($e^{\hat{\beta}-1.96SE}$, $e^{\hat{\beta}+1.96SE}$)	P-value
depc		-0.5004	0.6063	(0.4054, 0.9066)	0.0148
depc ²		-0.04073	0.9601	(0.9437, 0.9768)	<0.0001
agebase		0.03269	1.03351	(0.9970, 1.0714)	0.0750
agebase ²		-0.0004	0.99961	(0.9992, 1.0000)	0.0616
migration	1	-0.9619	0.3822	(0.1888, 0.7738)	0.0075
	2	0	1	.	.
familytype	1	0.01328	1.0134	(0.4841, 2.1212)	0.9719
	2	0.5215	1.6846	(0.8138, 3.4870)	0.1601
	3	0.7117	2.0375	(0.9654, 4.3002)	0.0618
	4	0	1	.	.
highqual	1	-0.6057	0.5457	(0.3276, 0.9089)	0.0200
	2	-0.4533	0.6355	(0.4335, 0.9317)	0.0202
	3	0	1	.	.
tenure	1	0.15	1.1618	(1.0954, 1.2323)	<0.0001
	2	0	1	.	.
casewave	four	0.3501	1.4192	(1.2057, 1.6706)	<0.0001
	two	0	1	.	.
see the key of variables on (page 87)					

members, number of dependent children, tenure status have more of an impact on $\log(\mathbf{ai})$ respectively, which means that the impact from family type on $\log(\mathbf{ai})$ depends on these three factors largely. The reason for the difference between the results from the EDA and the final model is because in EDA the relationship between family type and $\log(\mathbf{ai})$ has been studied independently, which doesn't include any other main effects, two-factor interactions or three-factor interactions. However, when statistical modelling is used, all of the related factors have been combined to contribute to the changes in $\log(\mathbf{ai})$. The impact from family type may be masked by the impacts from other factors.

5.4.2 Impacts on $\log(\mathbf{ai})$ from the factors of interest: migration and casewave

To work out the impacts on family $\log(\mathbf{ai})$ from migration and casewave, we use the LSMEANS statement and DIFF option in the PROC MIXED model, which calculates the difference of Least Squares Means of $\log(\mathbf{ai})$ between specified categories using our final model, averaging over all other categorical predictors and taking the average values of continuous predictors. Details are explained on Page 83. Here we summarise the results from the table of 'Differences of Least Squares Means' in Appendix C (page 128) to find out the income changes over time for mover families and stayer families respectively, where average value of $\text{depc} = 0.53$ and average value of $\text{agebase} = 50.77$.

Table 5.4: Summary of Differences of Least Squares Means of $\log(\mathbf{ai})$ between mover and stayer families

	Estimates $\hat{\beta}$	Ratio $e^{\hat{\beta}}$	Confidence interval of ratio ($e^{\hat{\beta}-1.96SE}$, $e^{\hat{\beta}+1.96SE}$)	P-value
Wave 2 $E(Y_{mw2}) - E(Y_{sw2})$	-0.07605	0.9268	(0.8331, 1.0309)	0.0035
Wave 4 $E(Y_{mw4}) - E(Y_{sw4})$	-0.1596	0.8525	(0.7660, 0.9487)	0.1618

Table 5.5: Summary of Differences of Least Squares Means of $\log(\mathbf{ai})$ between wave 4 and wave 2

	Estimates $\hat{\beta}$	Ratio $e^{\hat{\beta}}$	Confidence interval of ratio ($e^{\hat{\beta}-1.96SE}$, $e^{\hat{\beta}+1.96SE}$)	P-value
Mover families $E(Y_{mw4}) - E(Y_{mw2})$	0.01416	1.0147	(0.9180, 1.1217)	0.7750
Stayer families $E(Y_{sw4}) - E(Y_{sw2})$	0.09815	1.1031	(1.0846, 1.1220)	<0.0001

Table 5.4 shows that in wave 2 family income for movers is 92.68% of that of stayer, and the gap increased in wave 4, in which family income for movers is 85.25% of stayer's family income. We note that the estimate for wave 2 is significant with a P-value of 0.0035, but is not significant in wave 4 with a P-value of 0.1618.

Table 5.5 shows that family income for both mover and stayer families increased in wave 4 when compared with wave 2, and the estimated result is significant for stayer families and insignificant for mover families based on their P-values. The annual family income increased by about 1.10 times for stayer families, and about 1.01 times for mover families. The ratio of stayer family is slightly higher than that of mover family, which means that stayer families experienced a faster growth than did the mover families. This widened the gap of family income between mover and stayer intact families in wave 4, which has been reflected in Table 5.5. The result differs from that presented in Figure 4.3 (page 52), where the difference in $\log(\mathbf{ai})$ from wave 2 to wave 4 between mover and stayer families became smaller after the mover's internal migration. It shows the trends of $\log(\mathbf{ai})$ changes for mover families and stayer families by calculating means of $\log(\mathbf{ai})$ from the survey data, while Table 5.5 and Table 5.4 compare the estimated means between mover and stayer families produced by the General Linear Mixed Model in which the main effects and interactions between effects of all factors have been taken into account. So the difference is what we expected.

5.4.3 Impacts on $\log(\text{ai})$ from migration and casewave for families of different ages

In the final model, we found that interaction between age factors (including agebase and agebase^2), migration status and wave have a significant relationship with $\log(\text{ai})$. Here we illustrate how these interactions impact on $\log(\text{ai})$ by using the same exponentiation method as explained on page 87. Table 5.6 shows the ratio of family incomes between wave 4 and wave 2 of families in different migration status by age points. The ratios are worked out by keeping agebase and agebase^2 constant and averaging over all other predictors when analysing for each age point, see table 'Differences of Least Squares Means on page 128.

Table 5.6: Summary of Differences of Least Squares Means of $\log(\text{ai})$ between waves 4 and 2, $(\log(E(\text{ai})|\text{wave 4}) - \log(E(\text{ai})|\text{wave 2}))$, by migration status and age point

Age	migration status	Estimates $\hat{\beta}$	Ratio $e^{\hat{\beta}}$	Confidence interval of ratio $(e^{\hat{\beta}-1.96SE}, e^{\hat{\beta}+1.96SE})$	P-value
20	mover	0.6615	1.9377	(1.5866, 2.3665)	<0.0001
	stayer	0.1942	1.2143	(1.1430, 1.2902)	<0.0001
30	mover	0.2432	1.2753	(1.1398, 1.4269)	<0.0001
	stayer	0.1381	1.1481	(1.1116, 1.1858)	<0.0001
40	mover	-0.0398	0.9610	(0.8435, 1.0948)	0.5492
	stayer	0.0982	1.1032	(1.0784, 1.1286)	<0.0001
50	mover	-0.1878	0.8288	(0.7152, 0.9604)	0.0125
	stayer	0.0744	1.0722	(1.0512, 1.1040)	<0.0001
60	mover	-0.2005	0.8183	(0.7105, 0.9425)	0.0055
	stayer	0.0688	1.0691	(1.0428, 1.0960)	<0.0001
70	mover	-0.0782	0.9248	(0.7784, 1.0987)	0.3741
	stayer	0.0754	1.0783	(1.0491, 1.1083)	<0.0001

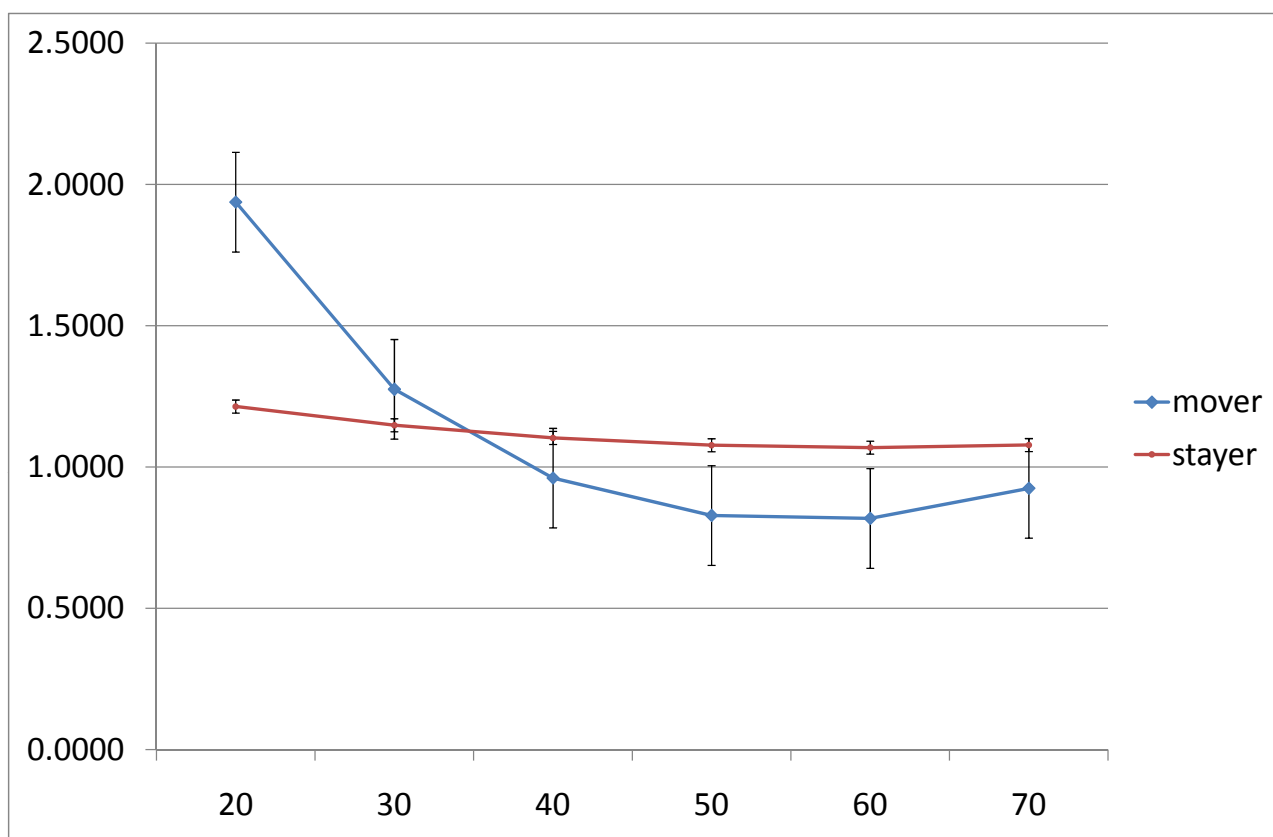
In Table 5.6, the Ratio column is determined by

$$e^{\hat{\beta}} = \frac{E(\text{ai}|\text{wave 4})}{E(\text{ai}|\text{wave 2})}$$

From this column can see that families of a younger average age, for example 20 and 30 year olds, benefit from the internal migration. For example,

for a family in which the average age of the adults is 20 years, $\log(\text{ai})$ is increased by about 1.94 times for mover families and about 1.21 times for stayer families from wave 2 to wave 4, which is the highest increase rate among all the age points in both mover and stayer families respectively. Then the rate of increase in annual income decreased while the age increased, see Figure 5.2 (page 94). However, mover families with an older average age of adult members experienced a decrease in their family annual incomes immediately after their movement. For instance, one year after their movement mover families aged over 50 years, only reached 82.88% of their income one year before their movement, and this percentage dropped even more as the average age of a family increased to 60 (some of the drops are not significant based on their P-values, such as age 40 and 70). While stayer families of an older average age have smaller increase rates than that of younger families, the rates are still positive which means their incomes increased between wave 2 and 4. This can be compared to the results from the EDA given in Figure 4.7 (page 58), which shows that $\log(\text{ai})$ is increased for all sample families which didn't take the migration status into account. Figure 4.8 (page 59) has shown the patterns of increase and decrease are similar with what we found from our final model. Therefore we believe that internal migration does impact on a family's income, increasing it for a younger family and decreasing it for an older family immediately.

Figure 5.2: Increase rate for family income between wave 4 and wave 2 for mover and stayer families by age point refer to Ratio column in Table 5.6



5.4.4 Estimating the difference in difference

From the previous section we know that the impact of migration status on family annual income depends on the average age of adult members of the family. We also discovered that the impact on family income of the interaction of migration status, wave and age are different for mover and stayer families. For example, in Table 5.6 (page 92) both mover and stayer families increase income from wave 2 to wave 4. We would like to know whether the amount of increase is significantly different between these two groups or not. Here we use a method called 'offset'. This method uses the difference from each observed value to a given value in our model instead of using an observed value. By using this method we can study the difference in $\log(ai)$ changes between mover and stayer families at each age point, in other words the difference in difference (Δ).

The Δ means the difference in changes of $\log(ai)$ over time between mover family and stayer families. The following method shows that we can use a single parameter directly from the model to test whether the changes are different between these two groups. To produce the estimates, then we 'offset' age impact by shifting age to the value we studying. For example, if we calculate the Δ for families in which the average age of adult members is 20 years, we need to create a new variable called *ages*, and let $ages = agebase - 20$. We then use this variable to replace *agebase* in the final model. For the model with *migration*, *ages*, $ages^2$, *casewave*, see the Equation 5.10.

$$\begin{aligned}
 E(Y) = E(\log(ai)) = & \beta_1 * ages + \beta_2 * ages^2 \\
 & + \beta_3 * migration + \beta_4 * migration * ages + \beta_5 * migration * ages^2 \\
 & + \beta_6 * casewave + \beta_7 * migration * ages * casewave \\
 & + \beta_8 * migration * ages^2 * casewave + \beta_9 * migration * casewave
 \end{aligned} \tag{5.10}$$

Let $migration = 1$ for mover family,

$migration = 0$ for stayer family,

$wave = 1$ for wave 4,

$wave = 0$ for wave 2,

then the difference of difference (Δ) at $age=20$ (i.e. $ages = 0$) becomes

$$\Delta = [E(Y_{mw4}) - E(Y_{mw2})] - [E(Y_{sw4}) - E(Y_{sw2})]$$

$$\begin{aligned}
&= [(\beta_3 + \beta_6 + \beta_9) - \beta_3] - [\beta_6 - 0] \\
&= \beta_9
\end{aligned}$$

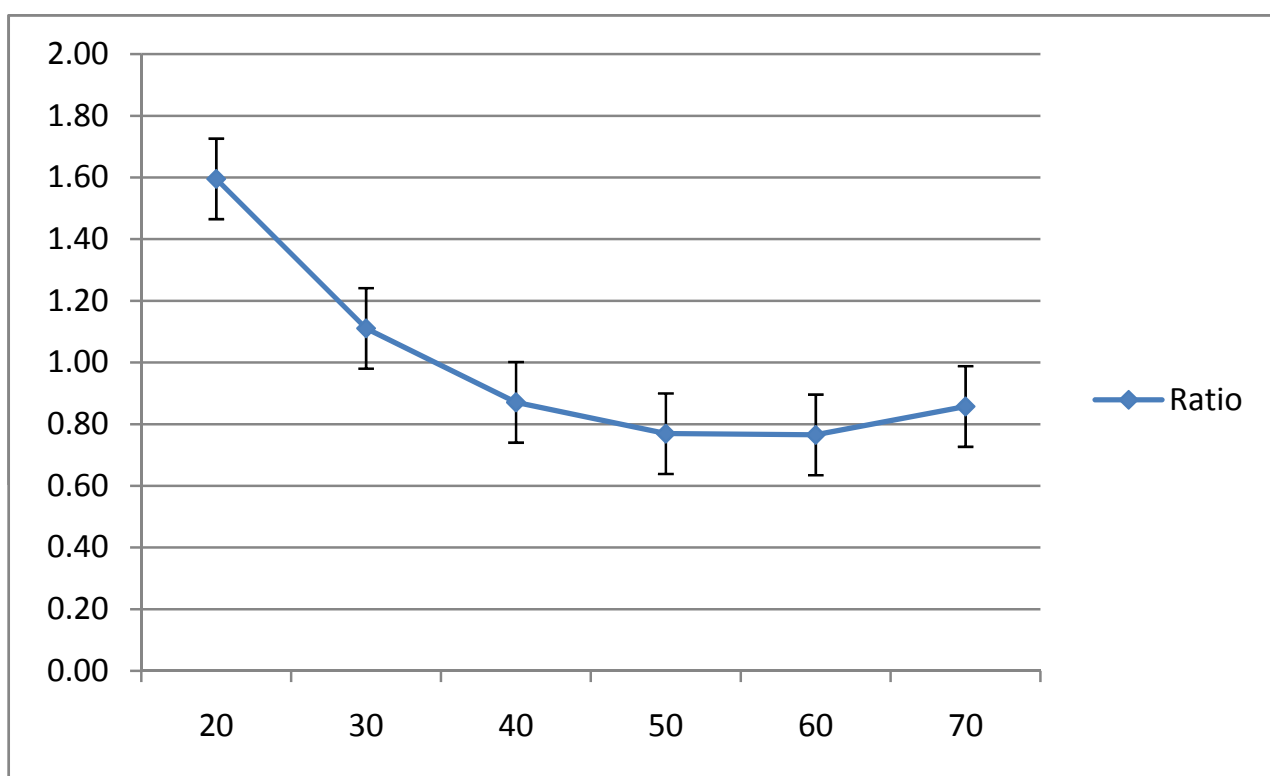
In another words, the estimate of Δ is the parameter estimate of interaction **migration * casewave**. For our final model, we replace **agebase** by **ages** and **agebase²** by **ages²**, then refit the model. The parameter associated with **migration * casewave** is Δ . This method can be used for any age bracket. For example, if we calculate Δ for a age of 40, we simply recalculate the **ages** and replace **agebase** and **agebase²** with **ages** and **ages²** in the final model of Equation 5.10 (page 95). The estimate of the parameter **migration * casewave** interaction is then the estimate of Δ for age 40. In Table 5.7 we summarise the estimates of Δ for selected age brackets.

Table 5.7: Summary of Differences in difference at each age points ($[E(Y_{mw4}) - E(Y_{mw2})] - [E(Y_{sw4}) - E(Y_{sw2})]$)

Age	Estimates $\hat{\beta}$	Ratio $e^{\hat{\beta}}$	Confidence interval of ratio ($e^{\hat{\beta}-1.96SE}, e^{\hat{\beta}+1.96SE}$)	P-value
20	0.4672	1.5955	(1.2962, 1.9640)	<0.0001
30	0.1051	1.1108	(0.9896, 1.2470)	0.0747
40	-0.1380	0.8711	(0.7625, 0.9951)	0.0403
50	-0.2621	0.7694	(0.6627, 0.8933)	0.0006
60	-0.2673	0.7654	(0.6632, 0.8834)	0.0003
70	-0.1536	0.8576	(0.7208, 1.0204)	0.0833

From Table 5.7 and Figure 5.3 (page 97)we can see that the estimated value of Δ is equivalent to the difference between the ration $e^{\hat{\beta}}$ between mover and stayer families at each age point in Table 5.6 (page92). The results show that internal migration brings greater benefit in a short time frame in terms of **log(ai)** growth for the families with a younger age. For example, in a family where the average age is 20 years, the increase in **log(ai)** for mover intact families is 1.5955 times the increase in stayer families over time (it shows significant impact based on the P-value). For families aged 40, 50, 60, and 70 years, the increase in **log(ai)** for mover families

Figure 5.3: Summary of Differences in difference in each age points refer to Ratio column in Table 5.7



is less than the increase in stayer families. For example, at age 60, the increase in mover families is 76.54% of the increase in stayer families, which is $0.8183/1.0691 = 0.7654$ in Table 5.6 (page 92). This means the gap of **log(ai)** between mover and stayer intact families has increased in families with average an age of 60 in wave 4.

Chapter 6

Conclusion and future work

In this thesis we have investigated the relationship between changes in a family's financial situation and internal migration. We have investigated short-term impacts on a family's financial situation using variables associated with family internal migration. Our analysis unit has been set as a family, because the internal migration affects entire families, even if it is sometimes driven by individuals. Impacts from internal migration on individuals will be reflected in the changes in the family financial situation. Our analysis unit is different from those used in the previous research mentioned in Section 1.1, therefore the results from our analysis is very difficult to compare with those previous results.

To construct a longitudinal family, we first defined the concept of internal migration on distance aspect and timing aspect. Based on this concept we identified movers and stayers in the SoFIE samples.

We then designed and compared four strategies to identify intact families. The chosen strategy used 'Families with the same adult members across waves' to define our longitudinal analysis unit- intact families.

After establishing all of the concepts, we set up our response variable to be the logarithm family annual income, and explored the relationship between this and the exploratory variables of family internal migration status, family type, average age of adult family members and its nonlinear form, number of dependent children and its quadratic form, the highest qualification of family members, dwelling tenure status, and case wave.

In EDA we found that mover intact families have lower **log(ai)** in all four waves when compared with stayer intact families. This indicates that mi-

gration status plays an important role in the changes to $\log(\mathbf{ai})$. We also found there are some three-factor interactions between internal migration, case wave and other factors relating to $\log(\mathbf{ai})$.

Following the EDA, we used the longitudinal analysis method of General Linear Mixed Model to analyse our data. We applied the method of variogram. We produced a variogram, and based on the variogram results, we set up the General Linear Mixed Model with the only random effect being the intercept. The models are implemented using the PROC MIXED procedure of SAS.

The results from our final model showed that the migration status as a main effect is not significant, and its impact on the logarithm of family annual income is reflected by the three-factor interaction between migration status, age and wave. This result indicated that the impacts from migration status on family $\log(\mathbf{ai})$ depend on wave and average age of adult members in the family. To infer the difference in impact between mover intact families and the control group stayer intact families over waves, we estimated contrasts between key subsets of the population: mover family vs. stayer families, change between mover families and the interaction between these effects.

Our final results indicate that internal migration does impact on the family annual income over time, and the impact depends on the average age of adult members of the family. Families with a younger average age benefit more from internal migration than do older families. For movers, older families aged 50 and 60 years experienced a loss in the family annual income and the youngest families aged 20 increased their families' income by about 1.9 times in the short-term after migration (mover families of age 40 and 70, the relationship between internal migration and family annual income is not significant over time). Stayer intact families increased in the family annual income at all age points. As stayer intact families have higher starting points in the family annual income and mover intact families have different increasing rate of different age point, for families in a young age group, the gap of the family annual income between mover and stayer intact families is reduced, whereas for older families the gap has been enlarged. From these results we conclude that families with a younger average age benefit from internal migration shortly after their movement, while families with an older average age experience a loss.

There is still the potential for future work to explore this topic further.

For example, with data from more waves of the SoFIE we could carry out the same study to assess the long-term impact of internal migration. The long-term estimation will determine if mover families experience income increase after a lag, and how long that lag is. We could also focus on other categorical response variables, such as labour-force involvement, to study the relationship between labour-force involvement and internal migration. In summary, internal migration is a interesting topic, not only for applying longitudinal analysis methodology, but because inference from this topic will be helpful for individuals, families, local governments and policy researchers.

Bibliography

2001 Census of Population and Dwellings: Incomes (2001).

URL: www.stats.govt.nz/census/2001-census-data/2001-census-incomes.aspx

A Longitudinal Survey of Income, Employment and Family Dynamics, Feasibility Project Final Report (2001).

URL: www.stats.govt.nz

Andrews, M., Clark, K. & Whittaker, W. (2007), 'The Employment and Earnings of Migrants in the UK'.

Area unit (2006).

URL: www.stats.govt.nz/surveys_and_methods/methods/classifications-and-standards/classification-related-stats-standards/area-unit/definition.aspx

Bill, A. & Mitchell, W. (2006), 'Great Expectations- migration and labour market outcomes in Australia', *Working Paper 06-08*.

Blackburn, M. L. (2006), 'The Impact of Internal Migration on Married Couples' Earnings in Britain, with a Comparison to the United States', *ISER Working Paper 2006-24*.

Diggle, P. J., Heagerty, P., Liang, K.-Y. & Zeger, S. L. (2002), *Analysis of Longitudinal Data*, second edn, Oxford University Press.

Dwelling definitions (2006).

URL: www.stats.govt.nz/census/about-2006-census/2006-census-definitions-questionnaires/definitions/dwelling.aspx

Family definitions (2006).

URL: www.stats.govt.nz/census/about-2006-census/2006-census-definitions-questionnaires/definitions/family.aspx

Household definitions (2006).

URL: www.stats.govt.nz/census/about-2006-census/2006-census-definitions-questionnaires/definitions/household.aspx

- Krieg, R. G. (1997), 'Occupational change, employer change, internal migration, and earnings', *Regional Science and Urban Economics* **27**, 1–15.
- Longitudinal Data Analysis with Discrete and Continuous Responses Course Notes* (2002).
URL: <https://support.sas.com>
- Meshblocks* (2001).
URL: www.stats.govt.nz/census/2001-census-data/definitions-and-questionnaires.aspx
- SAS/STAT(R) 9.22 User's Guide* (2010).
URL: <http://support.sas.com>
- Schutt, R. K. (2006), *Investigation the social world: the process and practice of research*, 5th, edn, Pine Forge Press.
- Technical Notes, Survey of Family, Income and Employment: Wave Four-September 2006* (2008).
URL: www.stats.govt.nz
- Territorial Authority* (2002).
URL: www.stats.govt.nz/surveys_and_methods/methods/classifications-and-standards/classification-related-stats-standards/territorial-authority/definition.aspx
- Yankow, J. J. (2003), 'Migration, job change, and wage growth: a new perspective on the pecuniary return', *Journal of Regional Science* **43-3**.

Appendix A

Keys for categorical variables in models

These keys have been used for all the model outputs in appendix.

Table A.1: Keys of categorical variables in models

Variables	Key	Description
migration	1	Mover family
	2	Stayer family
familytype	1	Not in a family nucleus
	2	One parent with child(ren)
	3	Couple only
	4	Couple with child(ren)
highqual	1	No qualification
	2	School or vocational qualification
	3	University qualification
tenure	1	own their residential property
	2	rent their residential property
casewave	one	wave one
	two	wave two
	three	wave three
	four	wave four

Appendix B

Initial model for analysis of wave 2 and 4 data

Figure B.1: SAS code for initial model, wave 2 and 4

```
proc mixed data=tfamai24;
  class id migration familytype tenure highqual casewave;
  model logai = depc sqdepc agebase sqagebase migration familytype
    highqual tenure casewave
    depc*agebase
    sqdepc*agebase
    depc*sqagebase
    sqdepc*sqagebase
    depc*migration
    sqdepc*migration
    agebase*migration
    sqagebase*migration
    depc*familytype
    sqdepc*familytype
    agebase*familytype
    sqagebase*familytype
    migration*familytype
    depc*highqual
    sqdepc*highqual
    agebase*highqual
    sqagebase*highqual
    migration*highqual
    familytype*highqual
    depc*tenure
    sqdepc*tenure
    agebase*tenure
    sqagebase*tenure
    migration*tenure
    familytype*tenure
    tenure*highqual
    depc*CaseWave
    sqdepc*CaseWave
    agebase*CaseWave
    sqagebase*CaseWave
    migration*CaseWave
    familytype*CaseWave
    highqual*CaseWave
    tenure*CaseWave
    depc*casewave*migration
    sqdepc*casewave*migration
    agebase*casewave*migration
    sqagebase*casewave*migration
    familytype*casewave*migration
    highqual*casewave*migration
    tenure*casewave*migration
    casewave*depc*agebase*migration
    casewave*sqdepc*sqagebase*migration
    casewave*sqdepc*agebase*migration
    casewave*depc*sqagebase*migration
    /solution ddfm=kr;
  random intercept/ subject=id ;
run;
```

Table B.1: Model Information, Dimensions, Iteration History

08:48 Thursday, October 14, 2010 1

The Mixed Procedure

Model Information	
Data Set	WORK.TFAMAI24
Dependent Variable	logai
Covariance Structure	Variance Components
Subject Effect	id
Estimation Method	REML
Residual Variance Method	Profile
Fixed Effects SE Method	Prasad-Rao-Jeske-Kackar-Harville
Degrees of Freedom Method	Kenward-Roger

Dimensions	
Covariance Parameters	2
Columns in X	208
Columns in Z Per Subject	1
Subjects	5712
Max Obs Per Subject	2
Observations Used	11313
Observations Not Used	151
Total Observations	11464

Iteration History			
Iteration	Evaluations	-2 Res Log Like	Criterion
0	1	23443.84665477	
1	2	21500.26268434	0.00000013
2	1	21500.26262671	0.00000000

Table B.2: Covariance Parameter Estimates, Fit Statistics

08:48 Thursday, October 14, 2010 2

The Mixed Procedure

Convergence criteria met.

Covariance Parameter Estimates		
Cov Parm	Subject	Estimate
Intercept	id	0.2408
Residual		0.2002

Fit Statistics	
-2 Res Log Likelihood	21500.3
AIC (smaller is better)	21504.3
AICC (smaller is better)	21504.3
BIC (smaller is better)	21517.6

Table B.3: Solution for Fixed Effects (continued to next page)

08:48 Thursday, October 14, 2010 3

The Mixed Procedure

Solution for Fixed Effects										
Effect	tenure	casewave	migration	familytype	highqual	Estimate	Standard Error	DF	t Value	Pr > t
Intercept						10.2319	0.4874	9421	20.99	<.0001
depc						-0.9078	0.4486	9813	-2.02	0.0430
sqdepc						0.06314	0.1249	9844	0.51	0.6133
agebase						0.02333	0.02206	9023	1.06	0.2901
sqagebase						-0.00025	0.000244	8588	-1.04	0.2988
migration			1			-1.2258	0.5715	1E4	-2.15	0.0320
migration			2			0
familytype				1		-0.05635	0.4640	9081	-0.12	0.9033
familytype				2		0.6278	0.3829	7079	1.64	0.1011
familytype				3		0.6623	0.4657	9611	1.42	0.1551
familytype				4		0
highqual					1	-0.6826	0.2872	7489	-2.38	0.0175
highqual					2	-0.4361	0.2085	7326	-2.09	0.0365
highqual					3	0
tenure	1					0.1130	0.1591	11E3	0.71	0.4776
tenure	2					0
casewave		four				0.3695	0.1058	5945	3.49	0.0005
casewave		two				0
depc*agebase						0.05400	0.02223	9865	2.43	0.0151
sqdepc*agebase						-0.00571	0.006849	1E4	-0.83	0.4044
depc*sqagebase						-0.00070	0.000277	1E4	-2.52	0.0118
sqdepc*sqagebase						0.000076	0.000096	11E3	0.80	0.4250
depc*migration			1			3.6873	4.2968	11E3	0.86	0.3908
depc*migration			2			0
sqdepc*migration			1			-0.6754	1.7037	11E3	-0.40	0.6918
sqdepc*migration			2			0

Table B.3: Solution for Fixed Effects (continued to next page)

08:48 Thursday, October 14, 2010 4

The Mixed Procedure

Solution for Fixed Effects										
Effect	tenure	casewave	migration	familytype	highqual	Estimate	Standard Error	DF	t Value	Pr > t
agebase*migration			1			0.03835	0.01828	8783	2.10	0.0360
agebase*migration			2			0
sqagebase*migration			1			-0.00036	0.000185	8708	-1.96	0.0498
sqagebase*migration			2			0
depc*familytype				1		0
depc*familytype				2		-0.1759	0.07853	8724	-2.24	0.0252
depc*familytype				3		0
depc*familytype				4		0
sqdepc*familytype				1		0
sqdepc*familytype				2		0.03926	0.01997	8725	1.97	0.0493
sqdepc*familytype				3		0
sqdepc*familytype				4		0
agebase*familytype				1		-0.03642	0.02122	8750	-1.72	0.0861
agebase*familytype				2		-0.05299	0.01874	7145	-2.83	0.0047
agebase*familytype				3		-0.02148	0.02129	9112	-1.01	0.3129
agebase*familytype				4		0
sqagebase*familytype				1		0.000348	0.000237	8363	1.47	0.1415
sqagebase*familytype				2		0.000639	0.000222	7260	2.88	0.0039
sqagebase*familytype				3		0.000138	0.000238	8591	0.58	0.5613
sqagebase*familytype				4		0
migration*familytype			1	1		0.09232	0.3749	11E3	0.25	0.8055
migration*familytype			1	2		-0.06322	0.2731	9827	-0.23	0.8170
migration*familytype			1	3		0.2669	0.3719	11E3	0.72	0.4730
migration*familytype			1	4		0
migration*familytype			2	1		0
migration*familytype			2	2		0

Table B.3: Solution for Fixed Effects (continued to next page)

08:48 Thursday, October 14, 2010 5

The Mixed Procedure

Solution for Fixed Effects										
Effect	tenure	casewave	migration	familytype	highqual	Estimate	Standard Error	DF	t Value	Pr > t
migration*familytype			2	3		0
migration*familytype			2	4		0
depc*highqual					1	-0.2130	0.1072	1E4	-1.99	0.0470
depc*highqual					2	-0.09345	0.05877	9408	-1.59	0.1118
depc*highqual					3	0
sqdepc*highqual					1	0.07176	0.02738	11E3	2.62	0.0088
sqdepc*highqual					2	0.03155	0.01115	11E3	2.83	0.0047
sqdepc*highqual					3	0
agebase*highqual					1	0.05139	0.01097	7075	4.68	<.0001
agebase*highqual					2	0.02843	0.007382	6996	3.85	0.0001
agebase*highqual					3	0
sqagebase*highqual					1	-0.00052	0.000106	6778	-4.88	<.0001
sqagebase*highqual					2	-0.00028	0.000065	6692	-4.28	<.0001
sqagebase*highqual					3	0
migration*highqual			1		1	0.1432	0.1647	9919	0.87	0.3847
migration*highqual			1		2	0.02243	0.1310	9744	0.17	0.8640
migration*highqual			1		3	0
migration*highqual			2		1	0
migration*highqual			2		2	0
migration*highqual			2		3	0
familytype*highqual				1	1	0.09525	0.1185	8481	0.80	0.4215
familytype*highqual				1	2	0.01658	0.08315	7120	0.20	0.8419
familytype*highqual				1	3	0
familytype*highqual				2	1	-0.1243	0.1163	8247	-1.07	0.2852
familytype*highqual				2	2	-0.1200	0.07858	7422	-1.53	0.1269
familytype*highqual				2	3	0

Table B.3: Solution for Fixed Effects (continued to next page)

08:48 Thursday, October 14, 2010 6

The Mixed Procedure

Solution for Fixed Effects										
Effect	tenure	casewave	migration	familytype	highqual	Estimate	Standard Error	DF	t Value	Pr > t
familytype*highqual				3	1	0.05432	0.1185	8843	0.46	0.6466
familytype*highqual				3	2	-0.05845	0.08174	7268	-0.72	0.4746
familytype*highqual				3	3	0
familytype*highqual				4	1	0
familytype*highqual				4	2	0
familytype*highqual				4	3	0
depc*tenure	1					-0.03187	0.05781	11E3	-0.55	0.5814
depc*tenure	2					0
sqdepc*tenure	1					0.003543	0.01251	11E3	0.28	0.7770
sqdepc*tenure	2					0
agebase*tenure	1					0.004629	0.005673	11E3	0.82	0.4146
agebase*tenure	2					0
sqagebase*tenure	1					-0.00008	0.000054	11E3	-1.54	0.1230
sqagebase*tenure	2					0
migration*tenure	1		1			0.003053	0.09683	9480	0.03	0.9748
migration*tenure	2		1			0
migration*tenure	1		2			0
migration*tenure	2		2			0
familytype*tenure	1			1		0.1318	0.07533	11E3	1.75	0.0803
familytype*tenure	2			1		0
familytype*tenure	1			2		-0.03130	0.06597	1E4	-0.47	0.6352
familytype*tenure	2			2		0
familytype*tenure	1			3		0.006296	0.07643	11E3	0.08	0.9344
familytype*tenure	2			3		0
familytype*tenure	1			4		0
familytype*tenure	2			4		0

Table B.3: Solution for Fixed Effects (continued to next page)

08:48 Thursday, October 14, 2010 7

The Mixed Procedure

Solution for Fixed Effects										
Effect	tenure	casewave	migration	familytype	highqual	Estimate	Standard Error	DF	t Value	Pr > t
tenure*highqual	1				1	0.04430	0.05481	11E3	0.81	0.4189
tenure*highqual	1				2	0.008354	0.03838	11E3	0.22	0.8277
tenure*highqual	1				3	0
tenure*highqual	2				1	0
tenure*highqual	2				2	0
tenure*highqual	2				3	0
depc*casewave		four				0.2648	0.3932	7285	0.67	0.5007
depc*casewave		two				0
sqdepc*casewave		four				-0.03215	0.1621	7617	-0.20	0.8427
sqdepc*casewave		two				0
agebase*casewave		four				-0.01101	0.003841	5858	-2.87	0.0042
agebase*casewave		two				0
sqagebase*casewave		four				0.000095	0.000035	5785	2.75	0.0061
sqagebase*casewave		two				0
migration*casewave		four	1			2.0152	0.5706	5855	3.53	0.0004
migration*casewave		two	1			0
migration*casewave		four	2			0
migration*casewave		two	2			0
familytype*casewave		four		1		-0.02759	0.04168	5891	-0.66	0.5080
familytype*casewave		two		1		0
familytype*casewave		four		2		-0.03119	0.03851	5841	-0.81	0.4180
familytype*casewave		two		2		0
familytype*casewave		four		3		-0.03138	0.04167	5903	-0.75	0.4515
familytype*casewave		two		3		0
familytype*casewave		four		4		0
familytype*casewave		two		4		0

Table B.3: Solution for Fixed Effects (continued to next page)

08:48 Thursday, October 14, 2010 8

The Mixed Procedure

Solution for Fixed Effects										
Effect	tenure	casewave	migration	familytype	highqual	Estimate	Standard Error	DF	t Value	Pr > t
highqual*casewave		four			1	0.04749	0.03003	5764	1.58	0.1138
highqual*casewave		two			1	0
highqual*casewave		four			2	0.01133	0.02066	5646	0.55	0.5835
highqual*casewave		two			2	0
highqual*casewave		four			3	0
highqual*casewave		two			3	0
tenure*casewave	1	four				0.02651	0.02285	6235	1.16	0.2461
tenure*casewave	1	two				0
tenure*casewave	2	four				0
tenure*casewave	2	two				0
depc*migrati*casewav		four	1			-3.9153	4.3439	6630	-0.90	0.3674
depc*migrati*casewav		two	1			0
depc*migrati*casewav		four	2			0
depc*migrati*casewav		two	2			0
sqdepc*migrat*casewa		four	1			0.8508	1.7118	6792	0.50	0.6192
sqdepc*migrat*casewa		two	1			0
sqdepc*migrat*casewa		four	2			0
sqdepc*migrat*casewa		two	2			0
agebas*migrat*casewa		four	1			-0.06034	0.01835	5828	-3.29	0.0010
agebas*migrat*casewa		two	1			0
agebas*migrat*casewa		four	2			0
agebas*migrat*casewa		two	2			0
sqageb*migrat*casewa		four	1			0.000517	0.000185	5820	2.80	0.0052
sqageb*migrat*casewa		two	1			0
sqageb*migrat*casewa		four	2			0
sqageb*migrat*casewa		two	2			0

Table B.3: Solution for Fixed Effects (continued to next page)

08:48 Thursday, October 14, 2010 9

The Mixed Procedure

Solution for Fixed Effects										
Effect	tenure	casewave	migration	familytype	highqual	Estimate	Standard Error	DF	t Value	Pr > t
migrat*family*casewa		four	1	1		-0.2161	0.3812	6053	-0.57	0.5708
migrat*family*casewa		two	1	1		0
migrat*family*casewa		four	1	2		-0.1417	0.2665	5957	-0.53	0.5951
migrat*family*casewa		two	1	2		0
migrat*family*casewa		four	1	3		-0.3734	0.3832	6070	-0.97	0.3298
migrat*family*casewa		two	1	3		0
migrat*family*casewa		four	1	4		0
migrat*family*casewa		two	1	4		0
migrat*family*casewa		four	2	1		0
migrat*family*casewa		two	2	1		0
migrat*family*casewa		four	2	2		0
migrat*family*casewa		two	2	2		0
migrat*family*casewa		four	2	3		0
migrat*family*casewa		two	2	3		0
migrat*family*casewa		four	2	4		0
migrat*family*casewa		two	2	4		0
migrat*highqu*casewa		four	1		1	-0.1466	0.1646	5734	-0.89	0.3731
migrat*highqu*casewa		two	1		1	0
migrat*highqu*casewa		four	1		2	-0.1320	0.1350	5782	-0.98	0.3279
migrat*highqu*casewa		two	1		2	0
migrat*highqu*casewa		four	1		3	0
migrat*highqu*casewa		two	1		3	0
migrat*highqu*casewa		four	2		1	0
migrat*highqu*casewa		two	2		1	0
migrat*highqu*casewa		four	2		2	0
migrat*highqu*casewa		two	2		2	0

Table B.3: Solution for Fixed Effects (end)

08:48 Thursday, October 14, 2010 10

The Mixed Procedure

Solution for Fixed Effects										
Effect	tenure	casewave	migration	familytype	highqual	Estimate	Standard Error	DF	t Value	Pr > t
migrat*highqu*casewa		four	2		3	0
migrat*highqu*casewa		two	2		3	0
migrat*tenure*casewa	1	four	1			-0.1545	0.1243	7182	-1.24	0.2141
migrat*tenure*casewa	1	two	1			0
migrat*tenure*casewa	2	four	1			0
migrat*tenure*casewa	2	two	1			0
migrat*tenure*casewa	1	four	2			0
migrat*tenure*casewa	1	two	2			0
migrat*tenure*casewa	2	four	2			0
migrat*tenure*casewa	2	two	2			0
depc*ageb*migr*casew		four	1			-0.03944	0.2224	11E3	-0.18	0.8593
depc*ageb*migr*casew		two	1			-0.1831	0.2595	11E3	-0.71	0.4804
depc*ageb*migr*casew		four	2			-0.01906	0.02068	7202	-0.92	0.3566
depc*ageb*migr*casew		two	2			0
sqde*sqag*migr*casew		four	1			-0.00018	0.001354	11E3	-0.13	0.8972
sqde*sqag*migr*casew		two	1			-0.00039	0.001520	11E3	-0.25	0.7994
sqde*sqag*migr*casew		four	2			-0.00004	0.000121	7209	-0.33	0.7407
sqde*sqag*migr*casew		two	2			0
sqde*ageb*migr*casew		four	1			0.003654	0.08760	11E3	0.04	0.9667
sqde*ageb*migr*casew		two	1			0.03231	0.1027	11E3	0.31	0.7532
sqde*ageb*migr*casew		four	2			0.002928	0.008909	7437	0.33	0.7424
sqde*ageb*migr*casew		two	2			0
depc*sqag*migr*casew		four	1			0.000943	0.003449	11E3	0.27	0.7845
depc*sqag*migr*casew		two	1			0.002345	0.003865	11E3	0.61	0.5440
depc*sqag*migr*casew		four	2			0.000283	0.000271	7066	1.04	0.2971
depc*sqag*migr*casew		two	2			0

Table B.4: Type 3 Tests of Fixed Effects (continued to next page)

08:48 Thursday, October 14, 2010 11

The Mixed Procedure

Type 3 Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
depc	1	9232	0.00	0.9453
sqdepc	1	9056	0.00	0.9732
agebase	1	7403	3.64	0.0564
sqagebase	1	7191	4.90	0.0268
migration	1	6294	0.45	0.5001
familytype	3	7618	4.70	0.0028
highqual	2	7726	3.14	0.0433
tenure	1	11E3	0.86	0.3536
casewave	1	5885	29.58	<.0001
depc*agebase	1	9009	0.00	0.9503
sqdepc*agebase	1	8944	0.01	0.9201
depc*sqagebase	1	8901	0.02	0.8998
sqdepc*sqagebase	1	8901	0.02	0.9024
depc*migration	1	9277	0.27	0.6038
sqdepc*migration	1	9112	0.04	0.8494
agebase*migration	1	5957	0.25	0.6164
sqagebase*migration	1	5897	0.40	0.5276
depc*familytype	1	8724	5.01	0.0252
sqdepc*familytype	1	8725	3.87	0.0493
agebase*familytype	3	7399	3.91	0.0084
sqagebase*familytype	3	7171	6.02	0.0004
migration*familytype	3	7010	0.40	0.7540
depc*highqual	2	9964	2.35	0.0959
sqdepc*highqual	2	11E3	5.88	0.0028
agebase*highqual	2	7221	12.35	<.0001
sqagebase*highqual	2	6853	14.64	<.0001

Table B.4: Type 3 Tests of Fixed Effects (continued to next page)

08:48 Thursday, October 14, 2010 12

The Mixed Procedure

Type 3 Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
migration*highqual	2	6914	0.52	0.5960
familytype*highqual	6	7683	1.18	0.3126
depc*tenure	1	11E3	0.30	0.5814
sqdepc*tenure	1	11E3	0.08	0.7770
agebase*tenure	1	11E3	0.67	0.4146
sqagebase*tenure	1	11E3	2.38	0.1230
migration*tenure	1	11E3	0.94	0.3323
familytype*tenure	3	11E3	3.89	0.0086
tenure*highqual	2	11E3	0.35	0.7032
depc*casewave	1	6635	0.61	0.4360
sqdepc*casewave	1	6800	0.21	0.6462
agebase*casewave	1	5847	20.10	<.0001
sqagebase*casewave	1	5835	14.63	0.0001
migration*casewave	1	5858	13.98	0.0002
familytype*casewave	3	5870	0.98	0.4026
highqual*casewave	2	5790	0.37	0.6933
tenure*casewave	1	7183	0.67	0.4145
depc*migrati*casewav	1	6630	0.81	0.3674
sqdepc*migrat*casewa	1	6792	0.25	0.6192
agebas*migrat*casewa	1	5828	10.82	0.0010
sqageb*migrat*casewa	1	5820	7.82	0.0052
migrat*family*casewa	3	5858	0.79	0.4992
migrat*highqu*casewa	2	5787	0.54	0.5853
migrat*tenure*casewa	1	7182	1.54	0.2141
depc*ageb*migr*casew	3	8758	0.44	0.7225
sqde*sqag*migr*casew	3	8853	0.06	0.9825

Table B.4: Type 3 Tests of Fixed Effects (end)

08:48 Thursday, October 14, 2010 13

The Mixed Procedure

Type 3 Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
sqde*ageb*migr*casew	3	8904	0.07	0.9757
depc*sqag*migr*casew	3	8730	0.47	0.7018

Appendix C

Final model for analysis of wave 2 and 4 data

Figure C.1: SAS code for final model, wave 2 and 4

SAS code for final model

```
proc mixed data=tfamai24;
  class id migration familytype tenure highqual casewave;
  model logai = depc sqdepc agebase sqagebase migration familytype
highqual tenure casewave
              depc*agebase
              depc*sqagebase
              agebase*migration
              sqagebase*migration
              depc*familytype
              sqdepc*familytype
              agebase*familytype
              sqagebase*familytype
              depc*highqual
              sqdepc*highqual
              agebase*highqual
              sqagebase*highqual
              familytype*tenure
              depc*CaseWave
              sqdepc*CaseWave
              agebase*CaseWave
              sqagebase*CaseWave
              migration*CaseWave
              agebase*casewave*migration
              sqagebase*casewave*migration
              /solution ddfm=kr;
  random intercept/ subject=id;
  lsmeans migration*casewave/ diff=all ;
  lsmeans migration*casewave/ diff=all at (agebase sqagebase)=(20
400);
  lsmeans migration*casewave/ diff=all at (agebase sqagebase)=(30
900);
  lsmeans migration*casewave/ diff=all at (agebase sqagebase)=(40
1600);
  lsmeans migration*casewave/ diff=all at (agebase sqagebase)=(50
2500);
  lsmeans migration*casewave/ diff=all at (agebase sqagebase)=(60
3600);
  lsmeans migration*casewave/ diff=all at (agebase sqagebase)=(70
4900);
run;
```

Table C.1: Model Information, Dimensions, Iteration History

09:09 Sunday, October 17, 2010 1

The Mixed Procedure

Model Information	
Data Set	WORK.TFAMAI24
Dependent Variable	logai
Covariance Structure	Variance Components
Subject Effect	id
Estimation Method	REML
Residual Variance Method	Profile
Fixed Effects SE Method	Prasad-Rao-Jeske-Kackar-Harville
Degrees of Freedom Method	Kenward-Roger

Dimensions	
Covariance Parameters	2
Columns in X	80
Columns in Z Per Subject	1
Subjects	5712
Max Obs Per Subject	2
Observations Used	11313
Observations Not Used	151
Total Observations	11464

Iteration History			
Iteration	Evaluations	-2 Res Log Like	Criterion
0	1	23219.93108603	
1	2	21248.68807974	0.00000004
2	1	21248.68806942	0.00000000

Table C.2: Covariance Parameter Estimates, Fit Statistics

09:09 Sunday, October 17, 2010 2

The Mixed Procedure

Convergence criteria met.

Covariance Parameter Estimates		
Cov Parm	Subject	Estimate
Intercept	id	0.2418
Residual		0.1999

Fit Statistics	
-2 Res Log Likelihood	21248.7
AIC (smaller is better)	21252.7
AICC (smaller is better)	21252.7
BIC (smaller is better)	21266.0

Table C.3: Solution for Fixed Effects (continued to next page)

09:09 Sunday, October 17, 2010 3

The Mixed Procedure

Solution for Fixed Effects										
Effect	tenure	casewave	migration	familytype	highqual	Estimate	Standard Error	DF	t Value	Pr > t
Intercept						10.1026	0.4003	9193	25.24	<.0001
depc						-0.5004	0.2053	9749	-2.44	0.0148
sqdepc						-0.04073	0.008795	11E3	-4.63	<.0001
agebase						0.03269	0.01836	8839	1.78	0.0750
sqagebase						-0.00039	0.000207	8543	-1.87	0.0616
migration			1			-0.9619	0.3599	8655	-2.67	0.0075
migration			2			0
familytype				1		0.01328	0.3769	9019	0.04	0.9719
familytype				2		0.5215	0.3712	7130	1.40	0.1601
familytype				3		0.7117	0.3811	9853	1.87	0.0618
familytype				4		0
highqual					1	-0.6057	0.2603	7467	-2.33	0.0200
highqual					2	-0.4533	0.1952	7438	-2.32	0.0202
highqual					3	0
tenure	1					0.1500	0.03303	11E3	4.54	<.0001
tenure	2					0
casewave		four				0.3510	0.08320	5837	4.22	<.0001
casewave		two				0
depc*agebase						0.03098	0.01049	9031	2.95	0.0031
depc*sqagebase						-0.00039	0.000134	8761	-2.90	0.0037
agebase*migration			1			0.04037	0.01663	8654	2.43	0.0152
agebase*migration			2			0
sqagebase*migration			1			-0.00040	0.000170	8653	-2.37	0.0177
sqagebase*migration			2			0
depc*familytype				1		0
depc*familytype				2		-0.2056	0.05690	9840	-3.61	0.0003

Table C.3: Solution for Fixed Effects (continued to next page)

09:09 Sunday, October 17, 2010 4

The Mixed Procedure

Solution for Fixed Effects										
Effect	tenure	casewave	migration	familytype	highqual	Estimate	Standard Error	DF	t Value	Pr > t
depc*familytype				3		0
depc*familytype				4		0
sqdepc*familytype				1		0
sqdepc*familytype				2		0.05094	0.01078	11E3	4.72	<.0001
sqdepc*familytype				3		0
sqdepc*familytype				4		0
agebase*familytype				1		-0.03953	0.01778	8679	-2.22	0.0262
agebase*familytype				2		-0.05356	0.01833	7259	-2.92	0.0035
agebase*familytype				3		-0.02500	0.01791	9217	-1.40	0.1629
agebase*familytype				4		0
sqagebase*familytype				1		0.000396	0.000203	8430	1.95	0.0513
sqagebase*familytype				2		0.000664	0.000217	7430	3.07	0.0022
sqagebase*familytype				3		0.000186	0.000204	8755	0.91	0.3614
sqagebase*familytype				4		0
depc*highqual					1	-0.3065	0.07028	9878	-4.36	<.0001
depc*highqual					2	-0.1023	0.04340	1E4	-2.36	0.0185
depc*highqual					3	0
sqdepc*highqual					1	0.09333	0.02179	11E3	4.28	<.0001
sqdepc*highqual					2	0.03307	0.009174	11E3	3.60	0.0003
sqdepc*highqual					3	0
agebase*highqual					1	0.05286	0.01054	7088	5.02	<.0001
agebase*highqual					2	0.02827	0.007109	7006	3.98	<.0001
agebase*highqual					3	0
sqagebase*highqual					1	-0.00053	0.000103	6808	-5.11	<.0001
sqagebase*highqual					2	-0.00027	0.000063	6691	-4.35	<.0001
sqagebase*highqual					3	0

Table C.3: Solution for Fixed Effects (continued to next page)

09:09 Sunday, October 17, 2010 5

The Mixed Procedure

Solution for Fixed Effects										
Effect	tenure	casewave	migration	familytype	highqual	Estimate	Standard Error	DF	t Value	Pr > t
familytype*tenure	1			1		0.08617	0.04193	11E3	2.06	0.0399
familytype*tenure	2			1		0
familytype*tenure	1			2		-0.05018	0.06231	1E4	-0.81	0.4207
familytype*tenure	2			2		0
familytype*tenure	1			3		-0.04059	0.04445	11E3	-0.91	0.3612
familytype*tenure	2			3		0
familytype*tenure	1			4		0
familytype*tenure	2			4		0
depc*casewave		four				-0.02905	0.02508	6622	-1.16	0.2468
depc*casewave		two				0
sqdepc*casewave		four				0.01547	0.007523	6888	2.06	0.0398
sqdepc*casewave		two				0
agebase*casewave		four				-0.00966	0.003235	5761	-2.98	0.0028
agebase*casewave		two				0
sqagebase*casewave		four				0.000081	0.000030	5716	2.69	0.0071
sqagebase*casewave		two				0
migration*casewave		four	1			1.5483	0.3475	5672	4.46	<.0001
migration*casewave		two	1			0
migration*casewave		four	2			0
migration*casewave		two	2			0
agebas*migrat*casewa		four	1			-0.06595	0.01616	5695	-4.08	<.0001
agebas*migrat*casewa		two	1			0
agebas*migrat*casewa		four	2			0
agebas*migrat*casewa		two	2			0
sqageb*migrat*casewa		four	1			0.000595	0.000166	5724	3.58	0.0003
sqageb*migrat*casewa		two	1			0

Table C.4: Solution for Fixed Effects (end); Type 3 Tests of Fixed Effects (continued to next page)

09:09 Sunday, October 17, 2010 6

The Mixed Procedure

Solution for Fixed Effects										
Effect	tenure	casewave	migration	familytype	highqual	Estimate	Standard Error	DF	t Value	Pr > t
sqageb*migrat*casewa		four	2			0
sqageb*migrat*casewa		two	2			0

Type 3 Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
depc	1	9667	13.07	0.0003
sqdepc	1	11E3	15.56	<.0001
agebase	1	6982	7.90	0.0050
sqagebase	1	6884	10.33	0.0013
migration	1	5710	0.35	0.5555
familytype	3	7677	4.83	0.0023
highqual	2	7792	3.32	0.0362
tenure	1	11E3	62.67	<.0001
casewave	1	5721	41.64	<.0001
depc*agebase	1	9031	8.73	0.0031
depc*sqagebase	1	8761	8.42	0.0037
agebase*migration	1	5735	0.25	0.6159
sqagebase*migration	1	5767	0.49	0.4830
depc*familytype	1	9840	13.06	0.0003
sqdepc*familytype	1	11E3	22.31	<.0001
agebase*familytype	3	7366	4.14	0.0061
sqagebase*familytype	3	7162	6.42	0.0002
depc*highqual	2	1E4	9.53	<.0001
sqdepc*highqual	2	11E3	12.06	<.0001
agebase*highqual	2	7260	13.78	<.0001

Table C.5: Type 3 Tests of Fixed Effects (end); Differences of Least Squares Means (continued to next page)

09:09 Sunday, October 17, 2010 7

The Mixed Procedure

Type 3 Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
sqagebase*highqual	2	6893	15.63	<.0001
familytype*tenure	3	11E3	4.20	0.0056
depc*casewave	1	6622	1.34	0.2468
sqdepc*casewave	1	6888	4.23	0.0398
agebase*casewave	1	5722	27.88	<.0001
sqagebase*casewave	1	5742	20.73	<.0001
migration*casewave	1	5672	19.86	<.0001
agebas*migrat*casewa	1	5695	16.65	<.0001
sqageb*migrat*casewa	1	5724	12.80	0.0003

Differences of Least Squares Means													
Effect	casewave	migration	_casewave	_migration	depc	sqdepc	agebase	sqagebase	Estimate	Standard Error	DF	t Value	Pr > t
migration*casewave	four	1	two	1	0.53	1.25	50.77	2885.9	0.01461	0.05112	5581	0.29	0.7750
migration*casewave	four	1	four	2	0.53	1.25	50.77	2885.9	-0.1596	0.05456	8723	-2.92	0.0035
migration*casewave	four	1	two	2	0.53	1.25	50.77	2885.9	-0.06144	0.05455	8721	-1.13	0.2601
migration*casewave	two	1	four	2	0.53	1.25	50.77	2885.9	-0.1742	0.05438	8659	-3.20	0.0014
migration*casewave	two	1	two	2	0.53	1.25	50.77	2885.9	-0.07605	0.05435	8657	-1.40	0.1618
migration*casewave	four	2	two	2	0.53	1.25	50.77	2885.9	0.09815	0.008653	5717	11.34	<.0001
migration*casewave	four	1	two	1	0.53	1.25	20.00	400.00	0.6615	0.1020	5646	6.48	<.0001
migration*casewave	four	1	four	2	0.53	1.25	20.00	400.00	0.1514	0.1118	8736	1.35	0.1756
migration*casewave	four	1	two	2	0.53	1.25	20.00	400.00	0.3456	0.1121	8769	3.08	0.0021
migration*casewave	two	1	four	2	0.53	1.25	20.00	400.00	-0.5101	0.1115	8682	-4.57	<.0001
migration*casewave	two	1	two	2	0.53	1.25	20.00	400.00	-0.3158	0.1113	8660	-2.84	0.0046
migration*casewave	four	2	two	2	0.53	1.25	20.00	400.00	0.1942	0.03092	5901	6.28	<.0001
migration*casewave	four	1	two	1	0.53	1.25	30.00	900.00	0.2432	0.05727	5607	4.25	<.0001

Table C.5: Differences of Least Squares Means (continued to next page)

09:09 Sunday, October 17, 2010 8

The Mixed Procedure

Differences of Least Squares Means													
Effect	casewave	migration	_casewave	_migration	depc	sqdepc	agebase	sqagebase	Estimate	Standard Error	DF	t Value	Pr > t
migration*casewave	four	1	four	2	0.53	1.25	30.00	900.00	-0.00863	0.06222	8679	-0.14	0.8896
migration*casewave	four	1	two	2	0.53	1.25	30.00	900.00	0.1295	0.06252	8762	2.07	0.0384
migration*casewave	two	1	four	2	0.53	1.25	30.00	900.00	-0.2519	0.06249	8738	-4.03	<.0001
migration*casewave	two	1	two	2	0.53	1.25	30.00	900.00	-0.1138	0.06220	8665	-1.83	0.0674
migration*casewave	four	2	two	2	0.53	1.25	30.00	900.00	0.1381	0.01650	5912	8.37	<.0001
migration*casewave	four	1	two	1	0.53	1.25	40.00	1600.0	-0.03982	0.06648	5620	-0.60	0.5492
migration*casewave	four	1	four	2	0.53	1.25	40.00	1600.0	-0.1304	0.07077	8806	-1.84	0.0655
migration*casewave	four	1	two	2	0.53	1.25	40.00	1600.0	-0.03218	0.07086	8830	-0.45	0.6497
migration*casewave	two	1	four	2	0.53	1.25	40.00	1600.0	-0.09053	0.07029	8683	-1.29	0.1978
migration*casewave	two	1	two	2	0.53	1.25	40.00	1600.0	0.007640	0.07017	8659	0.11	0.9133
migration*casewave	four	2	two	2	0.53	1.25	40.00	1600.0	0.09817	0.01162	5709	8.45	<.0001
migration*casewave	four	1	two	1	0.53	1.25	50.00	2500.0	-0.1878	0.07516	5594	-2.50	0.0125
migration*casewave	four	1	four	2	0.53	1.25	50.00	2500.0	-0.2138	0.08013	8738	-2.67	0.0076
migration*casewave	four	1	two	2	0.53	1.25	50.00	2500.0	-0.1394	0.08012	8736	-1.74	0.0820
migration*casewave	two	1	four	2	0.53	1.25	50.00	2500.0	-0.02601	0.07979	8656	-0.33	0.7445
migration*casewave	two	1	two	2	0.53	1.25	50.00	2500.0	0.04839	0.07977	8657	0.61	0.5441
migration*casewave	four	2	two	2	0.53	1.25	50.00	2500.0	0.07440	0.01247	5643	5.97	<.0001
migration*casewave	four	1	two	1	0.53	1.25	60.00	3600.0	-0.2005	0.07213	5564	-2.78	0.0055
migration*casewave	four	1	four	2	0.53	1.25	60.00	3600.0	-0.2588	0.07699	8665	-3.36	0.0008
migration*casewave	four	1	two	2	0.53	1.25	60.00	3600.0	-0.1920	0.07702	8674	-2.49	0.0127
migration*casewave	two	1	four	2	0.53	1.25	60.00	3600.0	-0.05831	0.07702	8667	-0.76	0.4491
migration*casewave	two	1	two	2	0.53	1.25	60.00	3600.0	0.008491	0.07697	8657	0.11	0.9122
migration*casewave	four	2	two	2	0.53	1.25	60.00	3600.0	0.06680	0.01270	5687	5.26	<.0001
migration*casewave	four	1	two	1	0.53	1.25	70.00	4900.0	-0.07818	0.08794	5681	-0.89	0.3741
migration*casewave	four	1	four	2	0.53	1.25	70.00	4900.0	-0.2656	0.09308	9009	-2.85	0.0043
migration*casewave	four	1	two	2	0.53	1.25	70.00	4900.0	-0.1902	0.09327	9053	-2.04	0.0414

Table C.6: Differences of Least Squares Means, (end)

09:09 Sunday, October 17, 2010 9

The Mixed Procedure

Differences of Least Squares Means													
Effect	casewave	migration	_casewave	_migration	depc	sqdepc	agebase	sqagebase	Estimate	Standard Error	DF	t Value	Pr > t
migration*casewave	two	1	four	2	0.53	1.25	70.00	4900.0	-0.1874	0.09119	8702	-2.06	0.0399
migration*casewave	two	1	two	2	0.53	1.25	70.00	4900.0	-0.1121	0.09099	8655	-1.23	0.2182
migration*casewave	four	2	two	2	0.53	1.25	70.00	4900.0	0.07538	0.01404	5682	5.37	<.0001

Appendix D

Final model for analysis of wave 1 to 4 data

Figure D.1: SAS code for final model, wave 1 to 4

SAS code for final model

```
proc mixed data=tfamail4 ;  
  class id migration familytype tenure highqual casewave;  
  model logai = depc sqdepc agebase sqagebase migration familytype  
    highqual tenure casewave  
    depc*agebase  
    depc*sqagebase  
    agebase*migration  
    sqagebase*migration  
    depc*familytype  
    sqdepc*familytype  
    agebase*familytype  
    sqagebase*familytype  
    depc*highqual  
    sqdepc*highqual  
    agebase*highqual  
    sqagebase*highqual  
    familytype*tenure  
    agebase*CaseWave  
    sqagebase*CaseWave  
    migration*CaseWave  
    tenure*CaseWave  
    agebase*casewave*migration  
    sqagebase*casewave*migration  
    /solution ddfm=kr;  
  random intercept/ subject=id ;  
run;
```

Table D.1: Model Information, Dimensions, Iteration History

09:09 Sunday, October 17, 2010 1

The Mixed Procedure

Model Information	
Data Set	WORK.TFAMAI14
Dependent Variable	logai
Covariance Structure	Variance Components
Subject Effect	id
Estimation Method	REML
Residual Variance Method	Profile
Fixed Effects SE Method	Prasad-Rao-Jeske-Kackar-Harville
Degrees of Freedom Method	Kenward-Roger

Dimensions	
Covariance Parameters	2
Columns in X	102
Columns in Z Per Subject	1
Subjects	5716
Max Obs Per Subject	4
Observations Used	22611
Observations Not Used	317
Total Observations	22928

Iteration History			
Iteration	Evaluations	-2 Res Log Like	Criterion
0	1	47399.64904465	
1	2	39405.43744591	0.00003665
2	1	39405.39973279	0.00000000

Table D.2: Covariance Parameter Estimates, Fit Statistics

09:09 Sunday, October 17, 2010 2

The Mixed Procedure

Convergence criteria met.

Covariance Parameter Estimates		
Cov Parm	Subject	Estimate
Intercept	id	0.2630
Residual		0.2078

Fit Statistics	
-2 Res Log Likelihood	39405.4
AIC (smaller is better)	39409.4
AICC (smaller is better)	39409.4
BIC (smaller is better)	39422.7

Table D.3: Solution for Fixed Effects (continued to next page)

09:09 Sunday, October 17, 2010 3

The Mixed Procedure

Solution for Fixed Effects										
Effect	tenure	casewave	migration	familytype	highqual	Estimate	Standard Error	DF	t Value	Pr > t
Intercept						10.2454	0.3445	13E3	29.74	<.0001
depc						-0.4920	0.1682	16E3	-2.93	0.0034
sqdepc						-0.03793	0.007220	22E3	-5.25	<.0001
agebase						0.02467	0.01599	12E3	1.54	0.1228
sqagebase						-0.00027	0.000182	12E3	-1.49	0.1352
migration			1			-0.9107	0.3707	12E3	-2.46	0.0140
migration			2			0
familytype				1		-0.1625	0.3241	12E3	-0.50	0.6160
familytype				2		0.3495	0.3302	9186	1.06	0.2898
familytype				3		0.7164	0.3161	15E3	2.27	0.0235
familytype				4		0
highqual					1	-0.5119	0.2349	9394	-2.18	0.0294
highqual					2	-0.5752	0.1729	9457	-3.33	0.0009
highqual					3	0
tenure	1					0.08897	0.02833	22E3	3.14	0.0017
tenure	2					0
casewave		four				0.3773	0.07890	17E3	4.78	<.0001
casewave		one				-0.4722	0.07851	17E3	-6.01	<.0001
casewave		thre				0.1205	0.07902	17E3	1.52	0.1273
casewave		two				0
depc*agebase						0.03194	0.008666	14E3	3.69	0.0002
depc*sqagebase						-0.00041	0.000111	13E3	-3.73	0.0002
agebase*migration			1			0.03765	0.01714	12E3	2.20	0.0281
agebase*migration			2			0
sqagebase*migration			1			-0.00037	0.000175	12E3	-2.14	0.0325
sqagebase*migration			2			0

Table D.3: Solution for Fixed Effects (continued to next page)

09:09 Sunday, October 17, 2010 4

The Mixed Procedure

Solution for Fixed Effects										
Effect	tenure	casewave	migration	familytype	highqual	Estimate	Standard Error	DF	t Value	Pr > t
depc*familytype				1		0
depc*familytype				2		-0.1657	0.04711	16E3	-3.52	0.0004
depc*familytype				3		0
depc*familytype				4		0
sqdepc*familytype				1		0
sqdepc*familytype				2		0.04462	0.008221	22E3	5.43	<.0001
sqdepc*familytype				3		0
sqdepc*familytype				4		0
agebase*familytype				1		-0.02934	0.01545	12E3	-1.90	0.0576
agebase*familytype				2		-0.04568	0.01601	9783	-2.85	0.0043
agebase*familytype				3		-0.02073	0.01525	13E3	-1.36	0.1740
agebase*familytype				4		0
sqagebase*familytype				1		0.000269	0.000178	11E3	1.51	0.1303
sqagebase*familytype				2		0.000548	0.000186	11E3	2.95	0.0032
sqagebase*familytype				3		0.000099	0.000177	12E3	0.56	0.5748
sqagebase*familytype				4		0
depc*highqual					1	-0.2928	0.05923	17E3	-4.94	<.0001
depc*highqual					2	-0.1051	0.03602	16E3	-2.92	0.0035
depc*highqual					3	0
sqdepc*highqual					1	0.08838	0.01857	19E3	4.76	<.0001
sqdepc*highqual					2	0.03450	0.007195	23E3	4.79	<.0001
sqdepc*highqual					3	0
agebase*highqual					1	0.04832	0.009660	8435	5.00	<.0001
agebase*highqual					2	0.03314	0.006404	8352	5.17	<.0001
agebase*highqual					3	0
sqagebase*highqual					1	-0.00048	0.000096	7803	-4.99	<.0001

Table D.3: Solution for Fixed Effects (continued to next page)

09:09 Sunday, October 17, 2010 5

The Mixed Procedure

Solution for Fixed Effects										
Effect	tenure	casewave	migration	familytype	highqual	Estimate	Standard Error	DF	t Value	Pr > t
sqagebase*highqual					2	-0.00032	0.000058	7641	-5.52	<.0001
sqagebase*highqual					3	0
familytype*tenure	1			1		0.06343	0.03297	2E4	1.92	0.0544
familytype*tenure	2			1		0
familytype*tenure	1			2		-0.01789	0.05014	19E3	-0.36	0.7213
familytype*tenure	2			2		0
familytype*tenure	1			3		-0.01490	0.03352	22E3	-0.44	0.6566
familytype*tenure	2			3		0
familytype*tenure	1			4		0
familytype*tenure	2			4		0
agebase*casewave		four				-0.01045	0.003289	17E3	-3.18	0.0015
agebase*casewave		one				0.01590	0.003275	17E3	4.85	<.0001
agebase*casewave		thre				-0.00264	0.003288	17E3	-0.80	0.4213
agebase*casewave		two				0
sqagebase*casewave		four				0.000085	0.000031	17E3	2.75	0.0060
sqagebase*casewave		one				-0.00016	0.000031	17E3	-5.10	<.0001
sqagebase*casewave		thre				0.000020	0.000031	17E3	0.64	0.5242
sqagebase*casewave		two				0
migration*casewave		four	1			1.4769	0.3518	17E3	4.20	<.0001
migration*casewave		one	1			0.3039	0.3478	17E3	0.87	0.3823
migration*casewave		thre	1			0.4017	0.3566	17E3	1.13	0.2601
migration*casewave		two	1			0
migration*casewave		four	2			0
migration*casewave		one	2			0
migration*casewave		thre	2			0
migration*casewave		two	2			0

Table D.3: Solution for Fixed Effects (end)

09:09 Sunday, October 17, 2010 6

The Mixed Procedure

Solution for Fixed Effects										
Effect	tenure	casewave	migration	familytype	highqual	Estimate	Standard Error	DF	t Value	Pr > t
tenure*casewave	1	four				0.01420	0.02121	17E3	0.67	0.5033
tenure*casewave	1	one				0.05867	0.02085	17E3	2.81	0.0049
tenure*casewave	1	thre				-0.01102	0.02073	17E3	-0.53	0.5952
tenure*casewave	1	two				0
tenure*casewave	2	four				0
tenure*casewave	2	one				0
tenure*casewave	2	thre				0
tenure*casewave	2	two				0
agebas*migrat*casewa		four	1			-0.06287	0.01636	17E3	-3.84	0.0001
agebas*migrat*casewa		one	1			-0.01889	0.01607	17E3	-1.18	0.2396
agebas*migrat*casewa		thre	1			-0.01832	0.01642	17E3	-1.12	0.2647
agebas*migrat*casewa		two	1			0
agebas*migrat*casewa		four	2			0
agebas*migrat*casewa		one	2			0
agebas*migrat*casewa		thre	2			0
agebas*migrat*casewa		two	2			0
sqageb*migrat*casewa		four	1			0.000566	0.000168	17E3	3.36	0.0008
sqageb*migrat*casewa		one	1			0.000229	0.000164	17E3	1.40	0.1622
sqageb*migrat*casewa		thre	1			0.000176	0.000167	17E3	1.05	0.2919
sqageb*migrat*casewa		two	1			0
sqageb*migrat*casewa		four	2			0
sqageb*migrat*casewa		one	2			0
sqageb*migrat*casewa		thre	2			0
sqageb*migrat*casewa		two	2			0

Table D.4: Type 3 Tests of Fixed Effects (continued to next page)

09:09 Sunday, October 17, 2010 7

The Mixed Procedure

Type 3 Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
depc	1	16E3	17.31	<.0001
sqdepc	1	2E4	13.00	0.0003
agebase	1	7736	13.66	0.0002
sqagebase	1	7628	14.60	0.0001
migration	1	5669	1.43	0.2313
familytype	3	9753	9.71	<.0001
highqual	2	1E4	5.54	0.0040
tenure	1	19E3	55.26	<.0001
casewave	3	17E3	23.34	<.0001
depc*agebase	1	14E3	13.58	0.0002
depc*sqagebase	1	13E3	13.92	0.0002
agebase*migration	1	5657	0.80	0.3702
sqagebase*migration	1	5654	0.84	0.3603
depc*familytype	1	16E3	12.37	0.0004
sqdepc*familytype	1	22E3	29.46	<.0001
agebase*familytype	3	9200	3.11	0.0252
sqagebase*familytype	3	8901	5.40	0.0010
depc*highqual	2	17E3	12.32	<.0001
sqdepc*highqual	2	21E3	17.45	<.0001
agebase*highqual	2	8770	16.89	<.0001
sqagebase*highqual	2	7964	19.02	<.0001
familytype*tenure	3	2E4	2.61	0.0494
agebase*casewave	3	17E3	13.36	<.0001
sqagebase*casewave	3	17E3	9.32	<.0001
migration*casewave	3	17E3	6.61	0.0002
tenure*casewave	3	17E3	4.16	0.0059

Table D.4: Type 3 Tests of Fixed Effects (end)

09:09 Sunday, October 17, 2010 8

The Mixed Procedure

Type 3 Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
agebas*migrat*casewa	3	17E3	5.26	0.0013
sqageb*migrat*casewa	3	17E3	3.90	0.0084

Appendix E

**Estimation results from offset age
20 in final model wave 2 and 4**

Table E.1: Model Information, Dimensions

09:47 Tuesday, November 2, 2010 1

The Mixed Procedure

Model Information	
Data Set	WORK.TFAMAI24
Dependent Variable	logai
Covariance Structure	Variance Components
Subject Effect	id
Estimation Method	REML
Residual Variance Method	Profile
Fixed Effects SE Method	Prasad-Rao-Jeske-Kackar-Harville
Degrees of Freedom Method	Kenward-Roger

Dimensions	
Covariance Parameters	2
Columns in X	80
Columns in Z Per Subject	1
Subjects	5712
Max Obs Per Subject	2
Observations Used	11313
Observations Not Used	151
Total Observations	11464

Table E.2: Solution for Fixed Effects (continued to next page)

09:09 Sunday, October 17, 2010 1

The Mixed Procedure

Solution for Fixed Effects										
Effect	tenure	casewave	migration	familytype	highqual	Estimate	Standard Error	DF	t Value	Pr > t
Intercept						10.6016	0.1300	9600	81.53	<.0001
depc						-0.03614	0.06625	11E3	-0.55	0.5854
sqdepc						-0.04073	0.008795	11E3	-4.63	<.0001
ages						0.03269	0.01836	8839	1.78	0.0750
sqages						-0.00039	0.000207	8543	-1.87	0.0616
migration			1			-0.3158	0.1113	8660	-2.84	0.0046
migration			2			0
familytype				1		-0.6190	0.1150	9583	-5.38	<.0001
familytype				2		-0.2840	0.1188	7641	-2.39	0.0168
familytype				3		0.2864	0.1187	11E3	2.41	0.0158
familytype				4		0
highqual					1	0.2404	0.1014	7835	2.37	0.0178
highqual					2	0.002166	0.08442	7779	0.03	0.9795
highqual					3	0
tenure	1					0.1500	0.03303	11E3	4.54	<.0001
tenure	2					0
casewave		four				0.1903	0.03375	5919	5.64	<.0001
casewave		two				0
depc*ages						0.03098	0.01049	9031	2.95	0.0031
depc*sqages						-0.00039	0.000134	8761	-2.90	0.0037
ages*migration			1			0.04037	0.01663	8654	2.43	0.0152
ages*migration			2			0
sqages*migration			1			-0.00040	0.000170	8653	-2.37	0.0177
sqages*migration			2			0
depc*familytype				1		0
depc*familytype				2		-0.2056	0.05690	9840	-3.61	0.0003

Table E.2: Solution for Fixed Effects (continued to next page)

09:09 Sunday, October 17, 2010 2

The Mixed Procedure

Solution for Fixed Effects										
Effect	tenure	casewave	migration	familytype	highqual	Estimate	Standard Error	DF	t Value	Pr > t
depc*familytype				3		0
depc*familytype				4		0
sqdepc*familytype				1		0
sqdepc*familytype				2		0.05094	0.01078	11E3	4.72	<.0001
sqdepc*familytype				3		0
sqdepc*familytype				4		0
ages*familytype				1		-0.03953	0.01778	8679	-2.22	0.0262
ages*familytype				2		-0.05356	0.01833	7259	-2.92	0.0035
ages*familytype				3		-0.02500	0.01791	9217	-1.40	0.1629
ages*familytype				4		0
sqages*familytype				1		0.000396	0.000203	8430	1.95	0.0513
sqages*familytype				2		0.000664	0.000217	7430	3.07	0.0022
sqages*familytype				3		0.000186	0.000204	8755	0.91	0.3614
sqages*familytype				4		0
depc*highqual					1	-0.3065	0.07028	9878	-4.36	<.0001
depc*highqual					2	-0.1023	0.04340	1E4	-2.36	0.0185
depc*highqual					3	0
sqdepc*highqual					1	0.09333	0.02179	11E3	4.28	<.0001
sqdepc*highqual					2	0.03307	0.009174	11E3	3.60	0.0003
sqdepc*highqual					3	0
ages*highqual					1	0.05286	0.01054	7088	5.02	<.0001
ages*highqual					2	0.02827	0.007109	7006	3.98	<.0001
ages*highqual					3	0
sqages*highqual					1	-0.00053	0.000103	6808	-5.11	<.0001
sqages*highqual					2	-0.00027	0.000063	6691	-4.35	<.0001
sqages*highqual					3	0

Table E.2: Solution for Fixed Effects (continued to next page)

09:09 Sunday, October 17, 2010 3

The Mixed Procedure

Solution for Fixed Effects										
Effect	tenure	casewave	migration	familytype	highqual	Estimate	Standard Error	DF	t Value	Pr > t
familytype*tenure	1			1		0.08617	0.04193	11E3	2.06	0.0399
familytype*tenure	2			1		0
familytype*tenure	1			2		-0.05018	0.06231	1E4	-0.81	0.4207
familytype*tenure	2			2		0
familytype*tenure	1			3		-0.04059	0.04445	11E3	-0.91	0.3612
familytype*tenure	2			3		0
familytype*tenure	1			4		0
familytype*tenure	2			4		0
depc*casewave		four				-0.02905	0.02508	6622	-1.16	0.2468
depc*casewave		two				0
sqdepc*casewave		four				0.01547	0.007523	6888	2.06	0.0398
sqdepc*casewave		two				0
ages*casewave		four				-0.00966	0.003235	5761	-2.98	0.0028
ages*casewave		two				0
sqages*casewave		four				0.000081	0.000030	5716	2.69	0.0071
sqages*casewave		two				0
migration*casewave		four	1			0.4672	0.1060	5621	4.41	<.0001
migration*casewave		two	1			0
migration*casewave		four	2			0
migration*casewave		two	2			0
ages*migrati*casewav		four	1			-0.06595	0.01616	5695	-4.08	<.0001
ages*migrati*casewav		two	1			0
ages*migrati*casewav		four	2			0
ages*migrati*casewav		two	2			0
sqages*migrat*casewa		four	1			0.000595	0.000166	5724	3.58	0.0003
sqages*migrat*casewa		two	1			0

Table E.2: Solution for Fixed Effects (end)

09:09 Sunday, October 17, 2010 4

The Mixed Procedure

Solution for Fixed Effects										
Effect	tenure	casewave	migration	familytype	highqual	Estimate	Standard Error	DF	t Value	Pr > t
sqages*migrat*casewa		four	2			0
sqages*migrat*casewa		two	2			0